Machine learning assisted determination of linearly independent set of generalized molecular coordinates

Ostojić, Tea

University of Zagreb
FACULTY OF SCIENCE
Department of Chemistry

Tea Ostojić

MACHINE LEARNING ASSISTED DETERMINATION OF LINEARLY INDEPENDENT SET OF GENERALIZED MOLECULAR COORDINATES

Diploma Thesis

submitted to the Department of Chemistry,
University of Zagreb Faculty of Science
for the academic degree of Master in Chemistry

Zagreb, 2021

The research presented in this Diploma Thesis was performed at the Division of Physical Chemistry, Department of Chemistry, Faculty of Science, University of Zagreb under mentorship of Professor Dr. Tomica Hrenar.

The study was supported by the Croatian Science Foundation in the frame of the project "Activity and *in silico* guided design of bioactive small molecules" (ADESIRE) (Project No: IP-2016-06-3775).

Acknowledgments

I would like to express my deep gratitude to my mentor, Professor Dr. Tomica Hrenar, for helping me and guiding me through this thesis.

I wish to show my appreciation to my reviewers, Professor Dr. Željka Soldin and Asisstant Professor Dr. Ivan Kodrin, for their feedback and advice.

Finally, I would like to thank my family and friends for supporting me during my studies.

Tea Ostojić

Table of Contents

University of Zagreb
Faculty of Science
Department of Chemistry

Diploma Thesis

ABSTRACT

MACHINE LEARNING ASSISTED DETERMINATION OF LINEARLY INDEPENDENT SET OF GENERALIZED MOLECULAR COORDINATES

Tea Ostojić

Trajectory analysis can provide information on the complete conformational or configurational space of molecular systems and their reactivity. Numerical analysis of trajectories can be performed in any type of coordinate system, *e.g.* Cartesian coordinate system, internal or normal coordinates, *etc*. In some applications, various coordinate systems will hold redundant information due to the linear dependence. A set of molecular coordinates contains null-vectors that introduce numerical instabilities and noise in the processing. In the case of a larger number of defined coordinates, there is also a problem with storage and data processing.

In this work, the procedure for constructing generalized and linearly independent set of internal molecular coordinates will be evaluated. Generalized internal coordinates are chemically intuitive and can be defined in a number of ways, for example: distances, angles or dihedral angles. By applying machine learning to molecular dynamics trajectory, a complete and linearly dependent set of generalized internal coordinate distances will be reduced to the linearly independent set that still contains all relevant information about conformational or configurational space, reactivity and molecular motion. Given procedure will be applied to the molecule of (*R*)-cinchonidine.

(83 pages, 12 figures, 21 tables, 31 references, original in English)

Thesis deposited in Central Chemical Library, University of Zagreb Faculty of Science, Horvatovac 102a, Zagreb, Croatia and in Repository of the Faculty of Science, University of Zagreb

Keywords: cinchonidine, conformational analysis, generalized coordinates, machine learning, molecular dynamics

Mentor: Dr. Tomica Hrenar, Professor

Reviewers:

1. Dr. Tomica Hrenar, Professor
2. Dr. Ivan Kodrin, Assistant Professor
3. Dr. Željka Soldin, Professor

Substitute: Dr. Branimir Bertoša, Professor

Date of exam: 18th May 2021

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Kemijski odsjek

Diplomski rad

SAŽETAK

ODREĐIVANJE LINEARNO NEZAVISNOG SKUPA GENERALIZIRANIH MOLEKULARNIH KOORDINATA PRIMJENOM STROJNOG UČENJA

Tea Ostojić

Analiza trajektorije može pružiti informacije o potpunom konformacijskom ili konfiguracijskom prostoru molekularnih sustava i njihovoj reaktivnosti. Numerička analiza trajektorija može se provesti u bilo kojem koordinatnom sustavu, npr. Cartesiusov koordinatni sustav, sustav internih ili normalnih koordinata, itd. No, u nekim aplikacijama različiti koordinatni sustavi zadržavaju suvišne informacije zbog linearne zavisnosti. Taj skup molekularnih koordinata sadrži nul-vektore koji onda uzrokuju numeričke nestabilnosti i numerički šum. U slučaju većeg broja definiranih koordinata također postoji i problem s pohranom te obradom podataka.

U ovom će se radu razraditi postupak za kreiranje generaliziranog i linearno nezavisnog skupa internih koordinata. Generalizirane interne koordinate kemijski su intuitivne, a mogu se definirati na više načina, na primjer: međuatomne udaljenosti, kutovi ili diedarski kutovi. Primjenom strojnog učenja na trajektoriju molekularne dinamike, potpun i linearno zavisan skup generaliziranih internih koordinata međuatomnih udaljenosti reducirat će se na linearno nezavisan skup, ali skup koji još uvijek sadrži sve relevantne informacije o konformacijskom ili konfiguracijskom prostoru, reaktivnosti i gibanju molekula. Razrađeni postupak primijenit će se na molekulu (*R*)-cinhonidina.

(83 stranice, 12 slika, 21 tablica, 31 literaturni navod, jezik izvornika: engleski jezik)

Rad je pohranjen u Središnjoj kemijskoj knjižnici Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu, Horvatovac 102a, Zagreb i Repozitoriju Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu

Ključne riječi: cinhonidin, generalizirane koordinate, konformacijska analiza, molekularna dinamika, strojno učenje

Mentor: prof. dr. sc. Tomica Hrenar

Ocjenitelji:

1. prof. dr. sc. Tomica Hrenar
2. doc. dr. sc. Ivan Kodrin
3. prof. dr. sc. Željka Soldin

Zamjena: prof. dr. sc. Branimir Bertoša

Datum diplomskog ispita: 18. svibnja 2021.

PROŠIRENI SAŽETAK

Cinhonidin je derivat kinuklidina, a svrstava se u skupinu alkaloida. Alkaloidi su organski spojevi koji imaju atom dušika s bazičnim svojstvima u svojoj strukturi, a često se koriste u farmakologiji zbog dokazanih ljekovitih svojstava. Alkaloidi se dijele prema tipu heterocikličkog prstena u molekuli, a cinhonidin spada u skupinu kinolinskih alkaloida koji imaju kinolinsku jezgru. Osim kinolinske jezgre cinhonidin sadrži i kinuklidinsku jezgru. Cinhonidin se u prirodi može naći u kori i lišću biljaka *Cinhone officinalis* i *Cinhone calisaye*, koje su se koristile u narodnoj medicini još od 15. stoljeća, a u 17. stoljeću[I] cinhonidin i službeno pronalazi upotrebu u medicini. Ova skupina spojeva ima antiparazitska svojstva, a danas se izučava s ciljem razvoja lijekova za borbu protiv kroničnih opstruktivnih plućnih bolesti.[II]

Valna funkcija $\Psi(\boldsymbol{r},t)$ je funkcija koja sadrži sve informacije o promatranom sustavu, a ovisi o položaju svih čestica promatranog sustava i vremenu[III]. Temeljna jednadžba kvantne mehanike je Schrödingerova jednadžba koja opisuje djelovanje operatora ukupne energije na valnu funkciju. Vremenski ovisna Schrödingerova jednadžba glasi:

$$i\hbar \frac{\partial |\Psi(\boldsymbol{r},t)\rangle}{\partial t} = \hat{H}|\Psi(\boldsymbol{r},t)\rangle \tag{1}$$

i sadrži sve informacije o sustavu i njegovoj propagaciji kroz vrijeme. Parcijalnom integracijom vremenski ovisne Schrödingerove jednadžbe može se doći do vremenski neovisne Schrödingerove jednadžbe koja opisuje stacionarno stanje sustava:

$$\hat{H}\Psi(\boldsymbol{r}) = E\Psi(\boldsymbol{r}) \tag{2}$$

u kojoj $\hat{H}$ predstavlja operator ukupne energije, $\Psi(\mathbf{r})$ valnu funkciju koja ovisi samo o položajima čestica u sustavu, a E pripadajuću energiju koja je ujedno i vlastita vrijednost operatora ukupne energije koji se naziva hamiltonijan. Operator ukupne energije može se rastaviti na operator kinetičke energije i operator potencijalne energije. Takva je jednadžba, osim za najjednostavnije sustave poput atoma vodika ili H_2^+, analitički nerješiva pa je za njezino

[I] J. Jaramillo-Arango, *Bot. J. Linn. Soc.* **53** (1949) 272–311.
[II] J.-P. Starck, L. Provins, B. Christophe, M. Gillard, S. Jadot, P. Lo Brutto, L. Que´re´, P. Talaga, M. Guyaux, *Bioorg. Med. Chem. Lett.* **18** (2008) 2675–2678.
[III] D. J. Griffiths, D. F. Schroeter, *Introduction to Quantum Mechanics*, Cambridge University Press, United Kingdom, 2018

rješavanje potrebno koristiti aproksimacije. Najčešće korištena aproksimacija je Born-Oppenheimerova aproksimacija, koja omogućava razdvajanje valne funkcije jezgri i valne funkcije elektrona promatranog sustava. Kako se masa elektrona i jezgara razlikuje za nekoliko redova veličine, posljedično se i brzine gibanja jezgara i elektrona značajno razlikuju. Elektroni, zbog manje mase u odnosu na masu jezgre, prema drugom Newtonovom zakonu, postižu puno veću akceleraciju i brzinu, a jezgre se gibaju znatno sporije u odnosu na njih. Elektroni se stoga praktički trenutno prilagođavaju maloj promjeni položaja jezgara nekog sustava, odnosno elektronska valna funkcija postaje parametarski ovisna o položajima jezgara. To omogućuje da se valna funkcija ne rješava simultano za sve čestice u sustavu nego se elektronska valna funkcija rješava za jedan položaj jezgara koji se u idućem koraku mijenja, a postupak rješavanja valne funkcije ponavlja se za svaku pojedinu konfiguraciju jezgara. Rezultat ovakvog postupka je ploha potencijalne elektronske energije.

Ploha potencijalne energije (engl. *Potential energy surface*, PES) funkcija je potencijalne elektronske energije u ovisnosti o geometrijskom opisu položaja jezgara[IV]. Ako su konformacije opisane Kartezijevim koordinatama, ploha potencijalne energije je $3N$-dimenzionalna funkcija, gdje N predstavlja broj jezgara u molekuli, a za njen bi prikaz bila potrebna $(3N+1)$-dimenzionalna hiperploha. Ako se konformacijska ploha razapinje internim koordinatama, ploha je $(3N-6)$-dimenzionalna funkcija za nelinearne sustave. Za njezin prikaz potrebna je $(3N-5)$-dimenzionalna hiperploha. U praksi je takve plohe gotovo nemoguće vizualizirati. Neke od metoda koje se koriste za pretragu konformacijskog prostora metode su sistematske pretrage, metode proizvoljnog pristupa i molekularna dinamika.

Molekularna dinamika skup je računalnih metoda koji se koristi za simulaciju sustava koji se sastoji od interagirajućih čestica i propagira kroz vrijeme s očuvanjem svojstava sustava[V]. Kao što je već spomenuto, rješavanje Schrödingerove jednadžbe za neki sustav je teško izvedivo, a ono se bez velikih pojednostavljenja može provesti samo na izrazito malim sustavima i tada se govori o kvantno-molekularnoj dinamici.

Klasična molekularna dinamika podrazumijeva pojednostavljenje gibanja jezgara sustava koristeći Newtonove jednadžbe za njihov opis i polje sila. Klasična molekularna dinamika stoga

[IV] A. R. Leach, *Molecular Modelling: Principles and Applications*, Pearson Education Limited, Dorchester, 2. izdanje, 2001.
[V] D. Marx, J. Hutter, *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*, Cambridge University Press, London, 2009.

slijedi Newtonove zakone gibanja. Stanje sustava definirano je u faznom prostoru 6*N* dimenzionalnosti, gdje *N* predstavlja broj čestica u nekom sustavu, a svaka je točka definirana položajem (tri prostorne koordinate *x, y, z*) i impulsima gibanja (u tri smjera p_x, p_y, p_z). U takvom prostoru definira se trajektoriju, odnosno funkciju koja predstavlja evoluciju vektora $\boldsymbol{X}$ koji sadrži informacije o svakoj čestici:

$$\boldsymbol{X} = (x_1, y_1, z_1, p_{x1}, p_{y1}, p_{z1}, x_2, y_2, z_2, p_{x2}, p_{y2}, p_{z2}, \dots, x_N, y_N, z_N, p_{xN}, p_{yN}, p_{zN}). \quad (3)$$

Semiklasična molekularna dinamika često se koristi u simulacijama smatanja proteina. Semiklasična molekularna dinamika također koristi Newtonove jednadžbe za opis gibanja jezgara, a iz potencijala se generiraju konzervativne sile iz kojih se numeričkom integracijom u određenom vremenskom intervalu računaju novi položaji i brzine jezgara. S obzirom na to da se radi o dinamici potrebno je krenuti od vremenski ovisne Schrödingerove jednadžbe i uvesti aproksimacije za njezino rješavanje. Kinetički dio hamiltonijana raspisuje se na kinetičku energiju jezgara i kinetičku energiju elektrona u sustavu, a potencijalni član na povoljne interakcije jezgara s elektronima i nepovoljne interakcije jezgara međusobno i elektrona međusobno:

$$\widehat{H} = \frac{1}{2m_e}\sum_{i=1}^{N}\nabla_i^2 + \frac{1}{2M_A}\sum_{A=1}^{M}\nabla_j^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{N}\frac{Z_A Z_B}{R_{AB}} \quad (4)$$

gdje *N* predstavlja broj elektrona u promatranom sustavu, *M* broj jezgara, M_A masu jezgre, m_e masu elektrona, a *r* i *R* udaljenosti između odabranih čestica. Uvođenjem aproksimacije zakočenih jezgara može se definirati tzv. elektronski hamiltonijan čiji izraz tada u atomskim jedinicama glasi:

$$\widehat{H}_{\text{el.}} = \frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{N}\frac{Z_A Z_B}{R_{AB}} \quad (5)$$

odnosno vrijedi:

$$\widehat{H} = \widehat{T}_N(\boldsymbol{R}) + \widehat{T}_e(\boldsymbol{r}) + \widehat{V}_{e,N}(\boldsymbol{r},\boldsymbol{R}) + \widehat{V}_{N,N}(\boldsymbol{R}) + \widehat{V}_{e,e}(\boldsymbol{r}) \quad (6)$$

gdje je $\boldsymbol{R}$ skup nuklearnih koordinata, a $\boldsymbol{r}$ skup elektronskih koordinata. Postojanje člana $\widehat{V}_{e,N}(\boldsymbol{r},\boldsymbol{R})$ ograničava da se u potpunosti separiraju jezgre od elektrona odnosno hamiltonijan prikazan jednadžbom (5) nije čisti elektronski hamiltonijan nego sadržava i tzv. mješoviti član koji opisuje interakcije elektrona i jezgara međusobno koji u realnim sustavima nije zanemariv. Već spomenuta Born-Oppenheimerova aproksimacija dopušta da se prostorna valna funkcija zapiše kao umnožak nuklearne valne funkcije i elektronske valne funkcije koja parametarski

ovisi o položaju jezgara $\Psi(\mathbf{r};\mathbf{R})\chi(\mathbf{R})$, fiksiramo li položaj jezgara u sustavu tzv. elektronski hamiltonijan i elektronsku Schrödingerovu jednadžbu može se zapisati kao:

$$\hat{H}_{\text{el.}} = \hat{T}_e(\boldsymbol{r}) + \hat{V}_{e,N}(\boldsymbol{r};\boldsymbol{R}) + \hat{V}_{N,N}(\boldsymbol{R}) + \hat{V}_{e,e}(\boldsymbol{r}) \tag{7}$$

$$\hat{H}_{\text{el.}}\Psi(\boldsymbol{r};\boldsymbol{R}) = E_{el.}\,\Psi(\boldsymbol{r};\boldsymbol{R}) \tag{8}$$

U aproksimaciji zamrznutih jezgara njihov položaj držimo konstantnim, potencijalni član $\hat{V}_{N,N}(\boldsymbol{R})$ također je konstantan i može se izostaviti iz elektronskog hamiltonijana:

$$\hat{H}_{\text{e.}} = \hat{T}_e(\boldsymbol{r}) + \hat{V}_{e,N}(\boldsymbol{r};\boldsymbol{R}) + \hat{V}_{e,e}(\boldsymbol{r}) \tag{9}$$

$$\hat{H}_{\text{e.}}\Psi(\boldsymbol{r};\boldsymbol{R}) = E_{\text{e.}}\,\Psi(\boldsymbol{r};\boldsymbol{R}) \tag{10}$$

Uz pretpostavku da je spektar elektronskog hamiltonijana diskretan, a da su njegovi svojstveni vektori ortogonalni, ukupnu valnu funkciju može se zapisati kao:

$$\Psi(\boldsymbol{r},\boldsymbol{R},t) = \sum_{l=1}^{\infty} \Psi_l\,(\boldsymbol{r};\boldsymbol{R})\chi_l(\boldsymbol{R},t) \tag{11}$$

gdje su $\chi_k(\boldsymbol{R},t)$ vremenski ovisni članovi ovakvog razvoja. Uvrštavanjem ovako zapisane ukupne valne funkcije u vremenski ovisnu Schrödingerovu jednadžbu, množenjem s lijeva $\Psi_k^*(\boldsymbol{R};\boldsymbol{r})$, integracijom po svim prostornim koordinatama elektrona $\boldsymbol{r}$ dobivaju se sljedeći izrazi:

$$i\hbar\frac{\partial\chi_k}{\partial t} = \left(-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 + E_k(\boldsymbol{R})\right)\chi_k + \sum_{l=1}^{\infty} C_{kl}\chi_l \tag{12}$$

$$C_{kl} = \int \Psi_k^*\left(-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2\right)\Psi_l d\boldsymbol{r} + \frac{1}{M_A}\left(\int \Psi_k^*(-\nabla_A)\Psi_l d\boldsymbol{r}\right)(-\nabla_A)\,, \tag{13}$$

C_{kl} predstavlja *kl*-ti element pravog neadijabatskog sprezanja. U adijabatskoj se aproksimaciji koriste samo dijagonalni elementi:

$$C_{kk} = \int \Psi_k^*\left(-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2\right)\Psi_l d\boldsymbol{r} \tag{14}$$

Niti jednom metodom ne može se u razvoju valne funkcije prema jednadžbi (14) uzeti beskonačan broj članova pa se ograničava na konačan broj realnih valnih funkcija, a radi pojednostavljenja zanemaruje se doprinos dijagonalnih članova C_{kk} iz čega slijedi:

$$i\hbar\frac{\partial\chi_k}{\partial t} = \left(-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 + E_k(\boldsymbol{R})\right)\chi_k \tag{15}$$

Valna funkcija χ_k opisuje se uvođenjem faktora amplitude A_k i faze S_k:

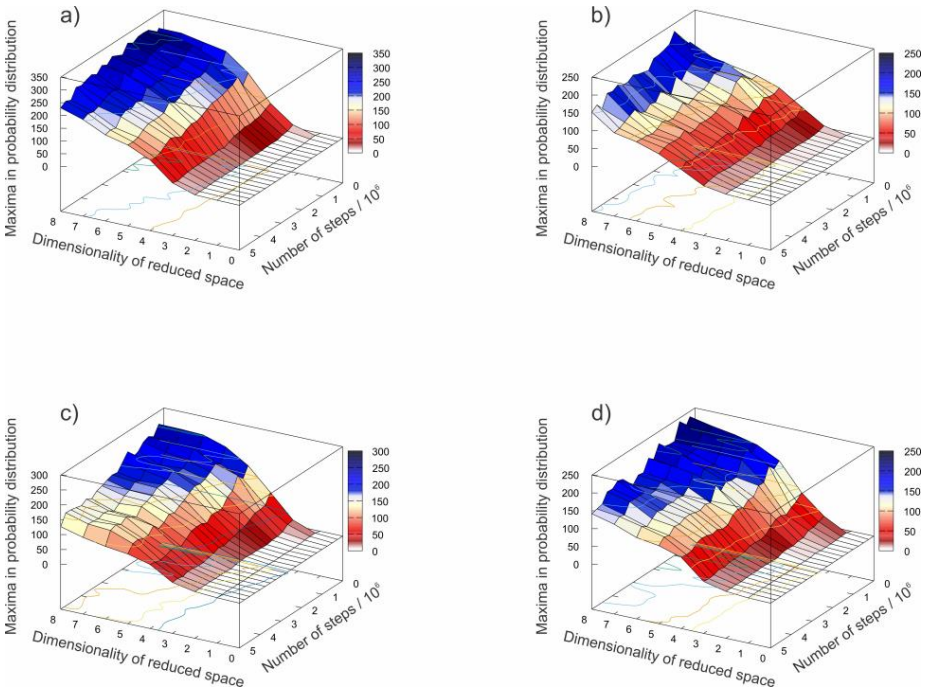
$$\chi_k(\boldsymbol{R}, t) = A_k(\boldsymbol{R}, t) e^{iS_k(\boldsymbol{R},t)} \quad (16)$$

Upravo se ovakav pristup, u kojem se vremenska ovisnost pripisuje samo klasičnom gibanju jezgara, a elektronska energija parametarski ovisi o njihovom položaju, koristi za reduciranje rješavanja vremenski ovisne Schrödingerove jednadžbe na rješavanje vremenski neovisne Schrödingerove jednadžbe. Ovakav se pristup još naziva i Born-Oppenheimerova molekularna dinamika (BOMD). Prelaskom na klasičnu sliku i povezivanjem s Newtonovim jednadžbama može se pokazati kako se jezgre gibaju u efektivnom potencijalu $V_{ef.}^{\mathrm{BO}}$ koji je opisan plohom potencijalne energiju u Born-Oppenheimerovoj aproksimaciji, a vremenski se neovisna Schrödingerova jednadžba računa kvantno-kemijskom metodom za fiksan položaj jezgara. Upravo se taj dobiveni potencijal za fiksne položaje jezgara koristi za generiranje novih položaja u idućem iteracijskom koraku. Potencijal koji se koristi za generiranje novih položaja jezgara u molekularnoj dinamici nije analitički poznat pa se generiranje novih položaja jezgara vrši numeričkom integracijom. Najprije je potrebno definirati integracijski korak, koji ne smije biti predug kako se ne bi narušila vjerodostojnost dinamike (primjerice ukoliko je duži od vremena prosječne vibracije rezultati molekularne dinamike, neće poslužiti za izvođenje zaključaka o svojstvima sustava) niti prekratak kako sama simulacija i prikupljanje podataka za analizu ne bi iziskivalo preveliko komputacijsko vrijeme. Postoji niz algoritama za numeričku integraciju, a u kemiji se koriste oni koji imaju vremensku reverzibilnost.

U ovom radu generirana je trajektorija (*R*)-cinhonidina od 5 000 000 točaka u Kartezijevim koordinatama (PM7 metoda, MOPAC2016[VI]) te interne koordinate, udaljenosti svih atoma međusobno za molekulu (*R*)-cinhonidina. Početne brzine dodijeljene su prema Maxwell-Boltzmannovoj raspodjeli pri temperaturi od 1273,15 K, a temperatura je držana konstantnom skaliranjem brzina. Integracijski korak iznosio je 0,5 fs, a ukupno trajanje iznosilo je 2,5 ps. Trajektorija je, korištenjem ***moonee***[VII] programskog koda, prevedena u trajektoriju u internim koordinatama. Pronađeni su svi lokalni maksimumi u raspodjeli gustoće vjerojatnosti koji odgovaraju minimumima na plohi potencijalne energije. Na trajektoriju s različitim brojem točaka, 1000, 2000, s korakom od 1000 do 10 000 točaka te s korakom od 10 000 do 100 000 točaka, primijenjen je algoritam strojnog učenja za dobivanje nezavisnog skupa generaliziranih

[VI] MOPAC2016, James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, HTTP://OpenMOPAC.net (2016)
[VII] T. Hrenar, ***moonee***, Code for Manipulation and Analysis of Multi- and Univariate Data, rev.

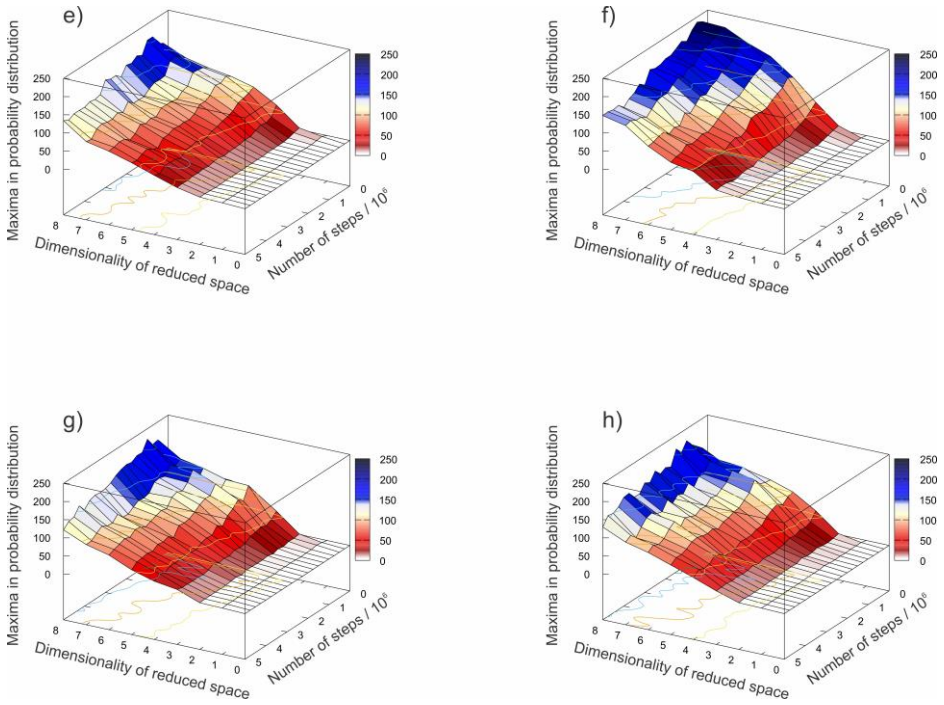
koordinata. Set internih molekularnih koordinata zapisan je u matričnom zapisu. U proceduri strojnog učenja korištena je QR dekompozicija za procjenu ranga matrice. Korištenjem metode izostavljanja jednog člana uzorka (*engl. leave-one-row-out*) ispitana je svaka interna koordinata. Ukoliko ona doprinosi rangu matrice, zadržana je, a ukoliko mu ne doprinosi, takva se uklanja iz matrice te je testirana njihova konvergencija. Konačan rezultat skup je koordinata koje su linearno nezavisne te se njima može opisati geometrija molekule bez gubitka značajnih svojstava ili redundantnih informacija. Osim skupa linearno nezavisnih koordinata praćeno je i vrijeme potrebno za izračun linearno nezavisnog seta koordinata. Za svaki skup linearno nezavisnih generaliziranih koordinata izračunati su platoi lokalnih maksimuma funkcije raspodjele vjerojatnosti u ovisnosti o broju koraka simulacije i dimenziji reduciranog prostora. Izračun strogih lokalnih maksimuma funkcije raspodjele vjerojatnosti proveden je prema već dobro uspostavljenom postupku objavljenom u prethodnom radu.[VIII]



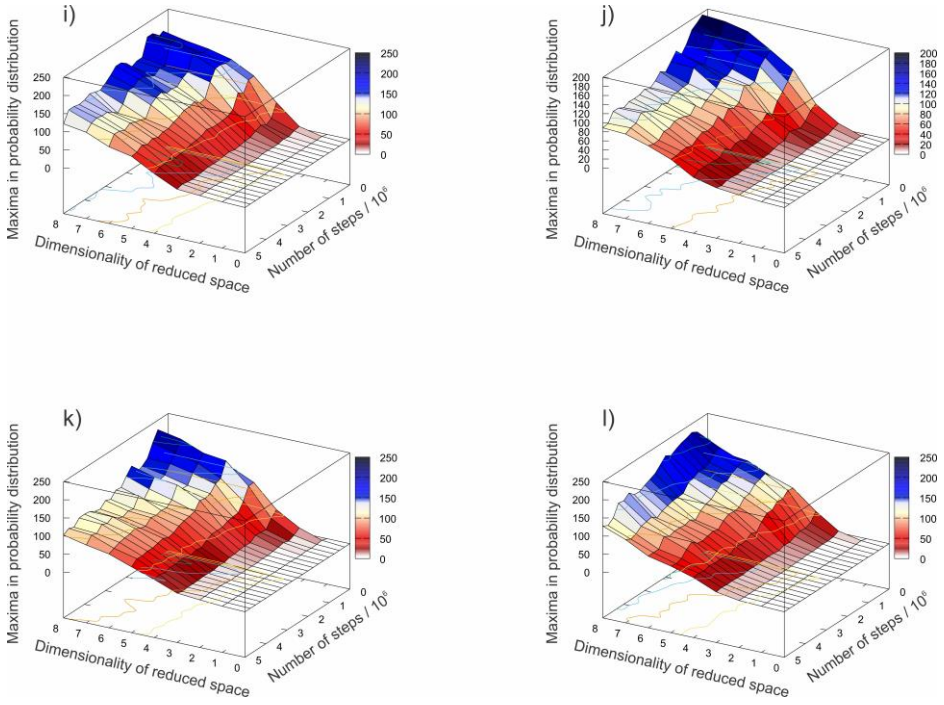
Slika I. Platoi lokalnih maksimuma u ovisnosti o ukupnom broju linearno nezavisnih unutarnjih koordinatnih udaljenosti za a) 1000, b) 2000, c) 3000, d) 4000 koraka simulacije.

(nastavlja se)

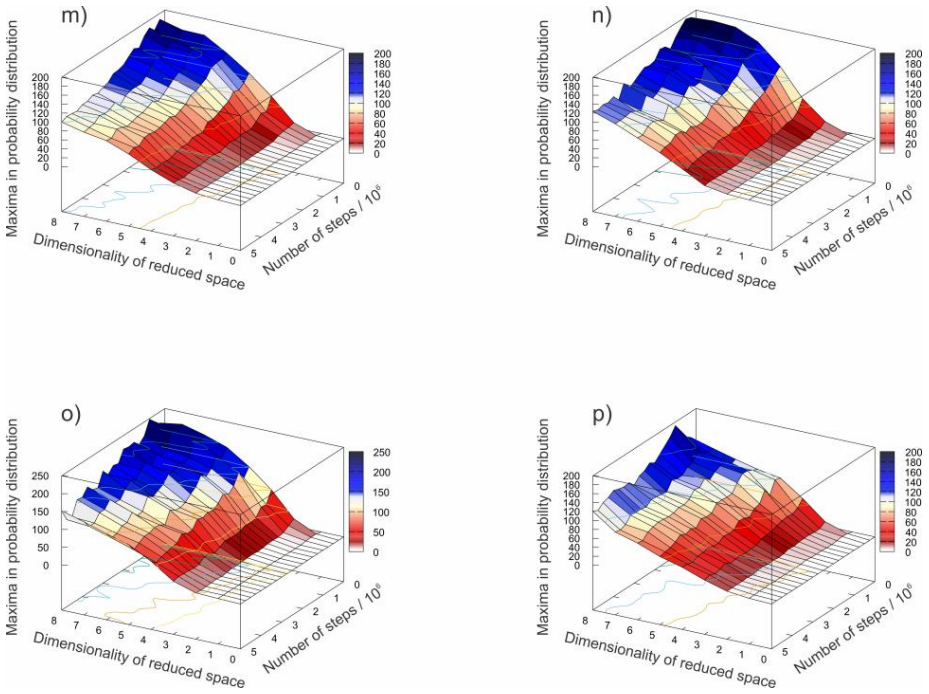
[VIII] K. Sović, T. Ostojić, S. Cepić, A. Ramić, R. Odžak, M. Skočibušić, T. Hrenar, I. Primožič, *Croat. Chem. Acta* **92** (2019) 259-267



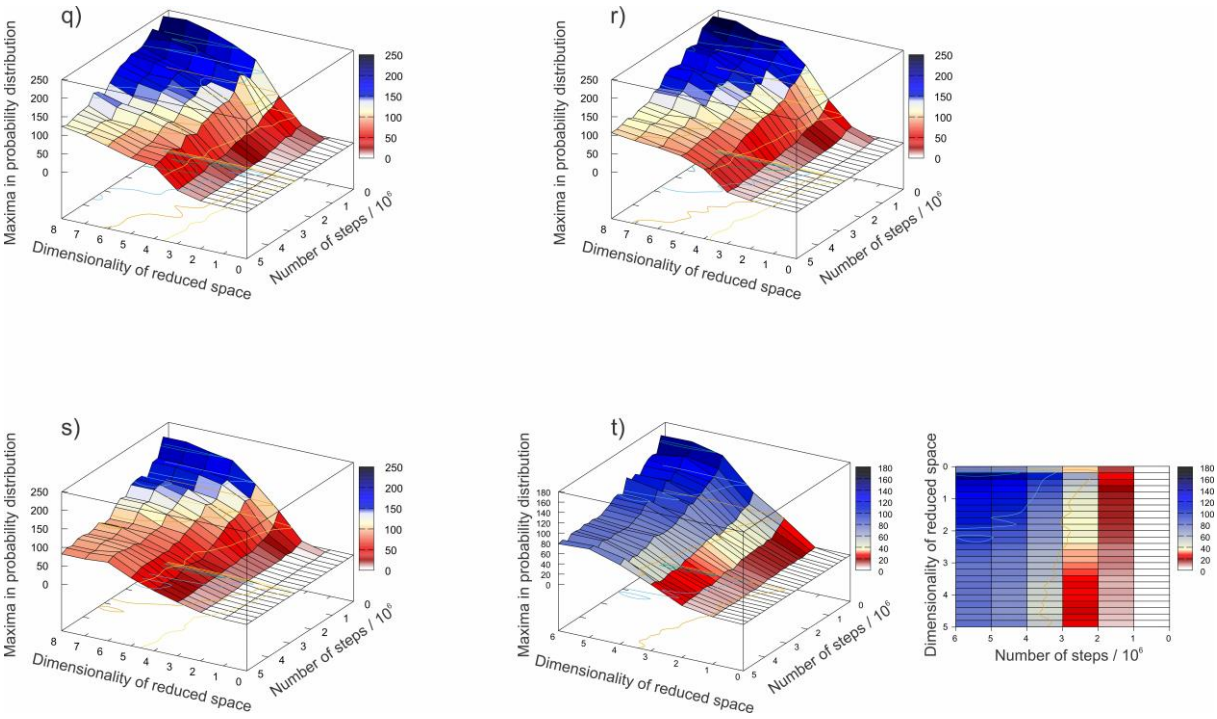
Slika I. Platoi lokalnih maksimuma u ovisnosti o ukupnom broju linearno nezavisnih unutarnjih koordinatnih udaljenosti za e) 5000, f) 6000, g) 7000, h) 8000 koraka simulacije. *(nastavlja se)*



Slika I. Platoi lokalnih maksimuma u ovisnosti o ukupnom broju linearno nezavisnih unutarnjih koordinatnih udaljenosti za i) 9000, j) 10 000, k) 20 000, l) 30 000 koraka simulacije. *(nastavlja se)*



Slika I. Platoi lokalnih maksimuma u ovisnosti o ukupnom broju linearno nezavisnih unutarnjih koordinatnih udaljenosti za m) 40 000, n) 50 000, o) 60 000, p) 70 000 koraka simulacije. *(nastavlja se)*



Slika I. Platoi lokalnih maksimuma u ovisnosti o ukupnom broju linearno nezavisnih unutarnjih koordinatnih udaljenosti za q) 80 000, r) 90 000, s) 100 000 koraka simulacije, t) odgovara prikazu iz prethodnog rada[VIII] za usporedbu.

Nakon 80 000 točaka simulacije broj internih koordinata udaljenosti konvergirao je, uz male promjene u definiciji (Appendix). Na to, osim prikaza na slici I., ukazuju i gradijenti koji su bili vrlo mali. Na slici I. q) – s) dobiveni su platoi koji pružaju jednaku raspodjelu vjerojatnosti i isti konformacijski prostor kao u prethodnom radu (slika I. t). Plato se nalazi na 6 glavnih komponenti i 2 000 000 točaka simulacije. Iako prethodni rad pokazuje da reducirani prostor oko platoa već sadrži sve informacije o konformacijskom prostoru, kao dodatna sigurnost pri određivanju istog simulacija se provodi s nešto većim brojem koraka.

§ 1. INTRODUCTION

The biggest challenge in molecular modelling is finding a certain way of figuring out the global minimum energy conformation. The distance geometry (DG) method was, firstly, used only for nuclear magnetic resonance (NMR) structural determination because it had big disadvantages in conformational analysis.[1] DG methods were introduced by Crippen and Havel in 1970s. DG method is a method that enables converting the distance ranges into a set in Cartesian coordinates. The molecular systems can be described as a set of minimum and maximum interatomic distances between all pairs of atoms.[1] The matrix defined by minimum and maximum interatomic distances contains the conformational space of a molecular system. The usual deterministic methods in conformational analysis have two big disadvantages. For example, in torsion search computational time grows exponentially with number of rotatable number of bonds in a molecule and granularity of a search. Another examples of distance geometry methods are the Monte Carlo method and molecular dynamics. In both, structures are generated from the first structure input.

On the other hand, genetic algorithm and some other geometry methods are randomly and independently generating the conformations in conformational space of a molecular system.[1] Usually, there are some limitations of the models. In these types of algorithms, every structure is scored and based on its score, it is or is not included in a final conformational space. Usually, structures are scored using force-fields or another evaluation based on energy of a molecule.

In 1997, D. C. Spellmeyer *et al.*[1] developed two distance geometry methods for conformational analysis. They showed that DG methods can be used in conformational sampling and that these methods can be as good or even better than the other conformational sampling methods that are commonly used. They are also, relatively, computationally inexpensive. One of the developed methods was distance geometry molecular dynamics that is close to reduced-coordinate molecular dynamics method. This method had disadvantage of requiring an input of parameters for every type of system that is investigated. Another method that was developed is based on random guess of initial structure in Cartesian coordinates by generating 4D coordinates for each atom and refinement using DG error function. As mentioned before, distance geometry is widely applied to NMR data. In this work the application of DG

methods is discussed, and it is concluded that they give comparable sampling that is quite simple and fast.

§ 2. LITERATURE REVIEW

2.1. Cinchonidine

Cinchonidine is a derivative of quinuclidine and is classified as a group of alkaloids. This molecule is of particular interest for us due to the ongoing scientific project and it was already thoroughly investigated in previous work. Alkaloids are organic compounds, which have a nitrogen atom with basic properties in their structure and are often used in pharmacology due to its medical properties. Alkaloids are divided according to the type of heterocyclic ring in molecules, and cinchonidine belongs to the group of quinoline alkaloids with a quinoline nucleus (Fig. 1).

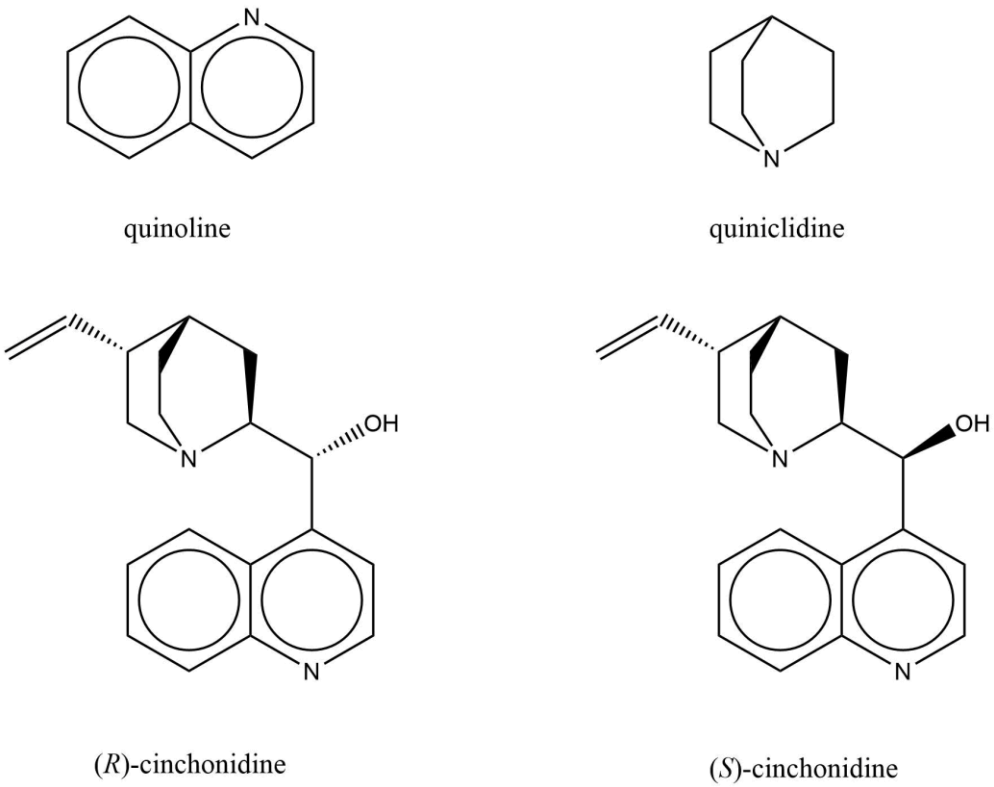


Figure 1. Structures of quinoline, quinuclidine, (*R*)- and (*S*)-cinchonidine.

Figure 1. presents structures of the aforementioned compounds. International Union of Pure and Applied Chemistry (IUPAC) names of shown compounds are: 1-benzopyridine, 1-azabicyclo[2.2.2]octane,

(*R*)−[(2*S*,4*S*,5*R*)5-ethenyl-1-azabicyclo[2.2.2]octan-2-yl](quinolin-4-yl)methanol and (*S*)−[(2*S*,4*S*,5*R*)-5-ethenyl-1-azabicyclo[2.2.2]octan-2-yl](quinolin-4-yl)methanol, respectively.

Cinchonidine can be found within the bark and leaves of the plants *Cinchona officinalis* and *Cinchona calisaya*, which have been used in alternative medicine since the 15th century, and later in the 17th century[2] as aid in curing sicknesses in hospitals. This group of compounds has antiparasitic properties and is currently being studied with the aim of developing drugs to cure chronic obstructive pulmonary disease (COPD)[3]. Until 1940, cinchonidine was used as antimalarial drug but had some serious side effects on people's health. Besides application in medicine, cinchonidine is used in organic synthesis as a starting material for the preparation of other quinuclidine derivatives. Especially, as a stationary phase for enantioselective chromatography and for directing chirality in syntheses.

2.2. Conformational Analysis of Cinchonidine

The conformational space of cinchonidine and its derivatives have been investigated by several methods in the past. Since the alkaloids from Cinchona have 5 stereocenters (Fig. 2), it is reasonable to expect variegated conformational behaviour.

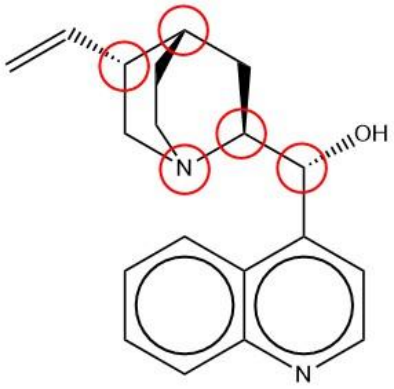


Figure 2. Cinchonidine's five stereocenters.

Some of these alkaloids are experimentally investigated using techniques such as Nuclear Magnetic Resonance Spectroscopy (NMR), Nuclear Overhauser Effect Spectroscopy (NOESY) in D_6-acetone. Beside these experiments, computational investigation is done. One of the first computational methods was done in 1989 by Dijkstra *et al.*[4] They have performed molecular mechanics (MM) calculations in vacuum with different force fields (MM2P and MMX). Later,

A. Vargas and A. Baiker[5] used relativistic Hamiltonian with spin-orbit coupling included to investigate conformational space of these alkaloids on a Pt111 surface. Baiker *et al.* reinvestigated conformational space of this group of compounds using NOESY and DFT metadynamic. In 2019, cinchonine and cinchonidine and their protonated and methylated quaternary derivatives were investigated using molecular dynamics by K. Sović *et al.* [6]

In 2008, Baiker *et al*.[5] used two different potentials for obtaining the potential energy surface of cinchonidine. They used BLYP functional with plane-wave basis sets using CPMD and B3PW91 functional with the Gaussian 6-311G(d,p) basis sets using Gaussian03. They obtained 11 conformers. The most stable conformer was conformer named Open(3).

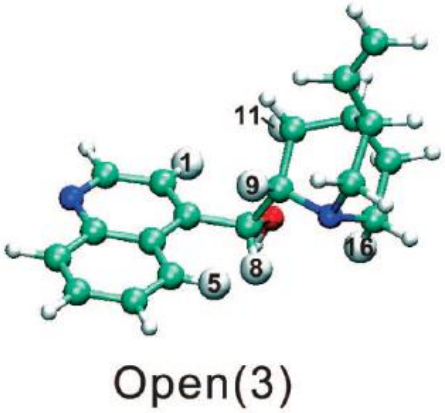


Figure 3. Cinchonidine conformation: Open(3).[5]

K. Sović *et al*. confirmed on high level of the theory, that the lowest energy conformer is the one labelled as Open(3). In mentioned work, simulations were performed using *on-the-fly* calculations of forces in each point of the simulation. All simulations were performed using PM7 Hamiltonian in MOPAC2016[7] using the program ***qcc***[8]. After obtaining the trajectory with 5 million steps, principal component analysis (PCA) was conducted with Nonlinear Iterative Partial Least Square (NIPALS) algorithm implemented in ***moonee***[9] to reduce the dimensionality. Probability distributions were generated in the reduced space and its maxima were taken as initial guesses for the optimization. Strict local maxima (SLM) of probability distribution indicates that the points around the molecule spend the most time in a simulation that is equal to strict local minima on potential energy surface (PES). All the maxima were optimized using B3LYP-D3 functional and 6-311++G(d,p) basis set with D3 version of Grimme's dispersion using Gaussian16[10] and clustered. The simulations and analysis were performed in Cartesian coordinate system. For the cinchonidine molecule that was used in this

work as well, namely as a test molecule, four conformers with the abundance over the 5% were obtained. Local strict maxima plateau was reached at approximately 2 million points in the simulations with first 4 principal components.

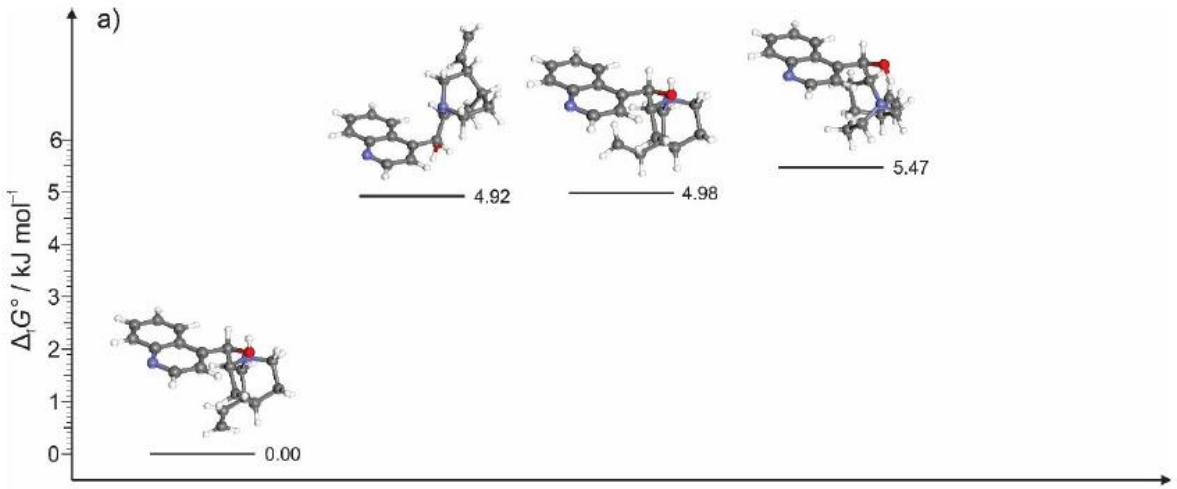


Figure 4. Four most stable conformers of cinchonidine with abundance of at least 5% at T = 298.15 K and p = 101325 Pa (B3LYP-D3/6-311++G(d,p) theory).[6]

Accordingly, in this work, new procedure for conformational analysis of Cinchona alkaloids using the generalized internal coordinate distances is developed.

§ 3. THEORETICAL SECTION

3.1. Conformational Analysis

Conformation is a 3-dimensional arrangement of atoms and groups of molecules that can be obtained by rotation around single (sigma) bonds. Chemical and physical properties of molecules are associated with their conformation. Potential energy surface (PES) is a function of potential electronic energy depending on the geometric description of the molecule.

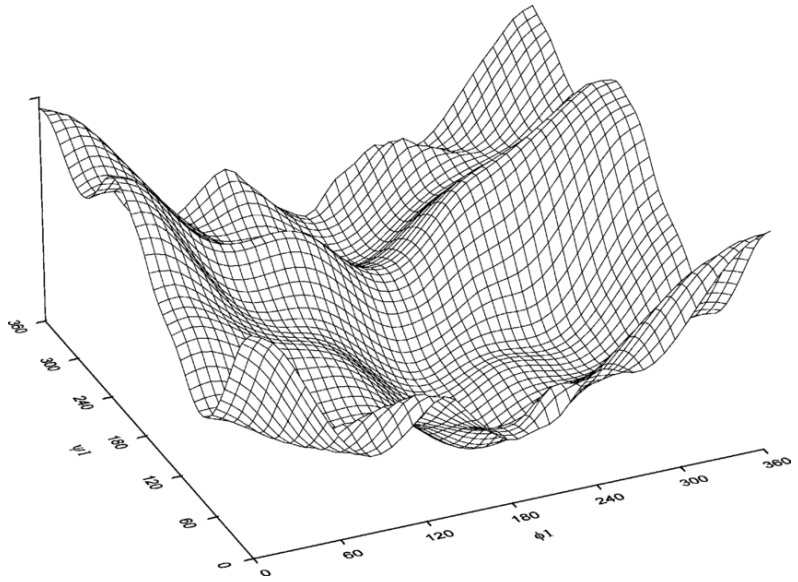


Figure 5. Example potential energy surface model.[11]

Figure 5. presents an example of potential energy surface, it presents the energy dependence on two torsion angles defined for bonds around which the rotation can occur. If the relative energy is dependent on only one parameter, it is known as one-dimensional function potential energy curve.

Conformational analysis shows high importance in analysing chemical reactions and active/binding sites of molecules where reactions occur. Different conformations of a system have different energies, often the minima of potential energy surface indicate the conformations of molecules that can be biologically active. In screening molecules for biological activity and their properties conformational analysis can significantly reduce the spent time. Usually, the goal of conformational analysis is to find a minima of potential energy surface and corresponding conformation, which is in that case called conformer. Some of the methods used

to search for conformational space are systematic search methods, random or stochastic methods and molecular dynamics.

Systematic search methods are methods by which the conformational space is searched within predictable changes in the conformation of a molecule. The most known systematic search method is *grid-search*, which is performed by systematically rotating atoms or groups around rotatable bonds for a certain amount of torsion angle. In this method, while generating new conformations, the bond length in the molecule and the angles between the bonds are kept constant. The disadvantage of this method is that the number of structures increases exponentially with increasing number of rotatable bonds in the molecule (so-called combinatorial explosion).[12] Each of the obtained conformations in this way can be subjected to optimization algorithm, but usually not all of them make sense.

Random methods are methods in which the new conformation in each step of calculation is generated randomly. The simplest algorithm is based on unsystematic change in Cartesian coordinates of atoms or the torsional angles of rotatable bonds. The obtained structure is minimized and taken as the initial structure for a new iteration during which a new conformation is again generated randomly. The iterative process is repeated for a certain number of steps or until it gives new conformations.

Molecular dynamics is a method used to simulate a system, propagating over time, of interacting particles. The trajectory is the result of a simulation of molecular dynamics and it is analysed to obtain the full conformational space of the observed system. The trajectory analysis of the system can be performed in any coordinate system, for example the Cartesian coordinate system. The use of a Cartesian coordinate system is not suitable for larger systems due to unnecessary information that causes problems with data processing and storage. If the conformations are described in a Cartesian coordinate system, the surface of potential energy is a $3N$-dimensional function where N represents the number of nuclei in the molecule and a $(3N + 1)$-dimensional hyperplane would be required to represent it. If the conformations are defined by internal coordinates, the surface is a $(3N – 6)$-dimensional function for nonlinear systems and a $(3N – 5)$-dimensional hyperplane would be required to represent it. In practice, hyperdimensional planes are impossible to visualize, so various methods are used, to generate conformations, that allow the dimensionality of the potential energy surface to be reduced. Such a set of coordinates contains null vectors; therefore, they are linearly dependent.

After obtaining the PES, all the minima are to be found. The minima can be local or global. Global minimum represents the conformer that has the lowest energy in the full conformational space, while the local minima represent all the other minima of PES function. Usually, the search is performed by finding critical points of the function and analysing the matrix of other derivatives at that critical point. The critical point is a point for which gradient of a function is null vector:

$$\nabla E(\boldsymbol{q_1}, \dots, \boldsymbol{q_{3N}}) = \left(\frac{\partial E}{\partial \boldsymbol{q_1}}, \dots, \frac{\partial E}{\partial \boldsymbol{q_{3N}}}\right) \tag{17}$$

After finding the critical point, the eigenvalues of the matrix of second derivatives, so-called Hessian matrix, are obtained by calculating second derivatives of the function at that point and the matrix is diagonalized. In a diagonalized matrix, diagonal elements represent eigenvalues of a matrix. If all eigenvalues are greater than zero, that point represents the minima of a function.

If some eigenvalues are negative, while other are positive, these points are called saddle points in which function has minima in one dimension, and maxima in others. Depending on the number of eigenvalues that are less than zero it is called n^{th} order saddle point. Molecular geometries in these points are called transition state structures and are especially important in the study of chemical reactions. They can indicate the mechanism of forming the certain species or conversion from one to another structure. This process is often difficult to carry out. The PES that we get from mentioned methods is not analytic function, therefore numerical methods are needed to find these points.

Another possibility is to do a statistical analysis of trajectories. The procedure for all local maxima in the probability distribution of the molecular geometry coordinate can be done. In molecular dynamics simulation it is expected that, during the simulation run, the molecule would statistically spend more time in and around the minima points on the PES. This assumption leads us to the conclusion that maxima points in the probability distribution of molecular geometry coordinates are equivalent to minima points in PES function. The advantage of this method is that it does not depend on the structure or connectivity of the molecule. It can be applied for the determination of the conformational space for the cyclic and noncyclic molecules without any postprocessing of the trajectory data.[13] Again, the problem is high dimensionality, and it should be reduced as much as possible without losing data important for interpretation. The most used method for that is principal component analysis (PCA) which is based on preserving most of the variance as in the initial data, where the variables are

correlated. The procedure is performed by calculating a vector that is a linear combination of vectors of correlated variables, in such a way that the variance for the processed data is maximal. The covariance matrix is calculated, and it gives the covariance between each pair of elements of a given vector. Diagonal elements of the matrix contain information on the variance of each variable separately, and on the non-diagonal elements are information on the covariance between two variables each other. The variance is maximized using LaGrange multipliers method which gives the vectors and its eigenvalues. These vectors represent principal components vectors. The greatest eigenvalue is associated with first principal component (PC1) and it describes the greatest percentage of variance. Usually, first 3 or 4 eigenvalues and their co-associated vectors (PC) are taken. All together, they describe enough of the variance in initial data so no essential information about system is lost.

Instead of using Cartesian coordinates, usually generalized coordinates are used. Generalized molecular coordinates are usually selected to provide the minimal number of independent coordinates that define the configuration of a system. For instance, they can be defined as changes in bond lengths, bond angles and torsion angles. For chemists, generalized coordinates are intuitive and often easier to use. However, in a set of molecular generalized coordinates, the null vectors are still present and coordinates are linearly dependent on each other.

In this work, a procedure for generating generalized molecular coordinates at each point of simulation of *ab initio* molecular dynamics will be developed. Machine learning algorithms will be applied to determine eigenvalues of the matrix of generalized molecular coordinates and eliminate linearly dependent molecular coordinates. The minimal set of generalized molecular coordinates that will contain all relevant information abound molecular motion will be obtained. In that way, repetitive information that cause problems mentioned above will not be present in the end. From the resulting set of generalized molecular coordinates, it is possible to get information about conformational or configurational space and reactivity of molecules.

3.2. Molecular Dynamics

Molecular dynamics is a set of computational methods used to simulate a system of interacting particles and propagate over time.[14,15] While simulating the system, properties of the system should be preserved. Solving the Schrödinger equation for a system is difficult to do. It can be carried out without large simplifications only on extremely simple and small systems, so it is possible to conduct quantum molecular dynamics. For different systems other types of

molecular dynamics are used. The group of methods that use Newton's equations to describe the motion of nuclei and use a force field to describe potentials at molecular level is called classical molecular dynamics. If instead of the force field, Schrödinger equation is calculated to generate the required potential in each step of the simulation, the group of methods is called semiclassical molecular dynamics. *Ab initio* molecular dynamics unifies approximate *ab initio* electronic structure theory and classical molecular dynamics.[14]

We should start from the time dependent Schrodinger equation and standard Hamiltonian[14]:

$$i\hbar\frac{\partial}{\partial t}\Psi(\{\boldsymbol{r}_i\},\{\boldsymbol{R}_A\},t) = \hat{H}\Psi(\{\boldsymbol{r}_i\},\{\boldsymbol{R}_A\},t) \tag{18}$$

$$\hat{H} = \frac{1}{2m_e}\sum_{i=1}^{N}\nabla_i^2 + \frac{1}{2M_A}\sum_{A=1}^{M}\nabla_j^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{N}\frac{Z_A Z_B}{R_{AB}} \tag{19}$$

for the electronic $\{\boldsymbol{r}_i\}$ and nuclear $\{\boldsymbol{R}_A\}$ degrees of freedom, in which N is the number of electrons in a system, M number of nuclei, M_A mass of nuclei, m_e mass of electron, r and R distance between particles.

Using the clamped nuclei approximation[16], the following equations for Hamiltonian in atomic units are obtained:

$$\hat{H} = \frac{1}{2m_e}\sum_{i=1}^{N}\nabla_i^2 + \hat{H}_{el} \tag{20}$$

$$\hat{H}_{\mathrm{el.}} = \frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{N}\frac{Z_A Z_B}{R_{AB}} \tag{21}$$

or in the other way:

$$\hat{H} = \hat{T}_N(\{\boldsymbol{R}\}) + \hat{T}_e(\{\boldsymbol{r}\}) + \hat{V}_{e,N}(\{\boldsymbol{r}\},\{\boldsymbol{R}\}) + \hat{V}_{N,N}(\{\boldsymbol{R}\}) + \hat{V}_{e,e}(\{\boldsymbol{r}\}) \tag{22}$$

in which $\boldsymbol{R}$ is the set of nuclear coordinates, and $\boldsymbol{r}$ is the set of electronic coordinates. The term $\hat{V}_{e,N}(\boldsymbol{r},\boldsymbol{R})$ does not allow us to separate electron from the nuclei. The Hamiltonian written as in equation (20) is not pure electronic Hamiltonian because of mixed term describing interactions between electrons and nuclei. In real systems that term cannot be ignored. Born-Oppenheimer approximation allows us to separate nuclei and electron[17], therefore, to write the spatial wave function as a product of the nuclear wave function and the electronic wave function which parametrically depends on the position of the nucleus. If we fix the position of the nucleus in the so-called electronic Hamiltonian, the electronic Schrödinger equation can be written as:

$$\hat{H}_{\text{el}} = \hat{T}_e(\{\boldsymbol{r}\}) + \hat{V}_{e,N}(\{\boldsymbol{r}\};\{\boldsymbol{R}\}) + \hat{V}_{N,N}(\{\boldsymbol{R}\}) + \hat{V}_{e,e}(\{\boldsymbol{r}\}) \tag{23}$$

$$\hat{H}_{\text{el}}\Psi(\{\boldsymbol{r}\};\{\boldsymbol{R}\}) = E_{el}\,\Psi(\{\boldsymbol{r}\};\{\boldsymbol{R}\}) \tag{24}$$

Term $\hat{V}_{N,N}(\boldsymbol{R})$ is constant so it can be excluded as following:[17]

$$\hat{H}_{\text{el}} = \hat{T}_e(\{\boldsymbol{r}\}) + \hat{V}_{e,N}(\{\boldsymbol{r}\};\{\boldsymbol{R}\}) + \hat{V}_{e,e}(\{\boldsymbol{r}\}) \tag{25}$$

$$\hat{H}_{\text{el}}\Psi(\{\boldsymbol{r}\};\{\boldsymbol{R}\}) = E_{\text{el}}\,\Psi(\{\boldsymbol{r}\};\{\boldsymbol{R}\}) \tag{26}$$

If we assume that the spectrum of Hamiltonian is discrete and the eigenfunctions are orthonormalized, we can write:[14]

$$\int \Psi_k^*(\{\boldsymbol{r}_i\}\,\{\boldsymbol{R}_A\})\Psi_l(\{\boldsymbol{r}_i\}\,\{\boldsymbol{R}_A\})\,d\boldsymbol{r} = \delta_{kl} \tag{27}$$

in which $d\boldsymbol{r}$ means integration over all i, all electron positions. Wave function can be written as:

$$\Psi(\{\boldsymbol{r}_i\}\,\{\boldsymbol{R}_A\},t) = \sum_{l=1}^{\infty}\Psi_l\,(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_A\})\chi_l(\{\boldsymbol{R}_A\},t) \tag{28}$$

The terms $\chi_l(\{\boldsymbol{R}_A\},t)$ can be observed as time-dependent expansion coefficients.

The next step is multiplication of the equation (28) with $\Psi_k^*(\{\boldsymbol{r}_i\}\,\{\boldsymbol{R}_A\})$ from the left side and integration over all electronic coordinates. The result is a set of coupled equations:[14]

$$i\hbar\frac{\partial\chi_k}{\partial t} = \left(-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 + E_k(\{\boldsymbol{R}_A\})\right)\chi_k + \sum_{l=1}^{\infty}C_{kl}\chi_l \tag{29}$$

$$C_{kl} = \int\Psi_k^*\left(-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2\right)\Psi_l d\boldsymbol{r} + \frac{1}{M_A}\left(\int\Psi_k^*(-\nabla_A)\Psi_l d\boldsymbol{r}\right)(-\nabla_A)\,, \tag{30}$$

in which C_{kl} is non-coupling operator. The first shown term is a matrix element from nuclei's kinetic energy operator and the second one depends on their momenta.[14] In adiabatic approximation only diagonal elements are used:

$$C_{kk} = \int\Psi_k^*\left(-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2\right)\Psi_l d\boldsymbol{r} \tag{31}$$

If the electronic function is real, the second term in equation (29) equals zero. That leads us to the decoupling from the set of coupled equations (29) and (30):[14]

$$i\hbar\frac{\partial\chi_k}{\partial t} = \left[-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 + E_k(\{\boldsymbol{R_A}\}) + C_{kk}(\{\boldsymbol{R_A}\})\right]\chi_k \tag{32}$$

In practice, additional approximations can be used, for example limiting the number of terms $\chi_l(\{\boldsymbol{R}_A\},t)$ in equation (27) and neglecting the C_{kk} terms, we get:

$$i\hbar\frac{\partial\chi_k}{\partial t} = \left[-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 + E_k(\{\boldsymbol{R_A}\})\right]\chi_k \tag{33}$$

The equation (32) is called Born-Oppenheimer approximation[16]. The following step in molecular dynamics is approximating nuclei as a classical part particle. That approximation will allow to use Newton's equations of motion. The Born-Oppenheimer approximation can be used in almost all physical situations, so it will be used for the following derivation. In smaller number of cases Born-Oppenheimer approximation is not valid, which are not going to be considered in this work.

To get semiclassical mechanics from quantum mechanics, the wave function should be written as in terms of an amplitude factor A_k and a phase S_k, which are both considered to be real and $A_k > 0$ in this polar representation:

$$\chi_k(\{\boldsymbol{R_A}\}, t) = A_k(\{\boldsymbol{R_A}\}, t)e^{iS_k(\boldsymbol{R},t)} \tag{34}$$

After inserting equation (34) to equation (33) and separating real and imaginary parts, the following equations are obtained:[14]

$$\frac{\partial S_k}{\partial t} + \sum_{A=1}^{M}\frac{1}{2M_A}(\nabla_A S_k)^2 + E_k = \hbar^2\sum_{A=1}^{M}\frac{1}{2M_A}\frac{\nabla_A^2 A_k}{A_k} \tag{35}$$

$$\frac{\partial A_k}{\partial t} + \sum_{A=1}^{M}\frac{1}{2M_A}(\nabla_A A_k)(\nabla_A S_k) + \sum_{A=1}^{M}\frac{1}{2M_A}A_k(\nabla_A S_k)^2 = 0 \tag{36}$$

The equation (36) is multiplied by A_k from the left:

$$\frac{\partial {A_k}^2}{\partial t} + \sum_{A=1}^{M}\frac{1}{2M_A}\left({A_k}^2\nabla_A S_k\right) = 0 \tag{37}$$

$$\frac{\partial \rho_k}{\partial t} + \sum_{A=1}^{M}\nabla_A \boldsymbol{J}_{k,A} = 0 \tag{38}$$

in which ρ_k is nuclear probability density, and $\boldsymbol{J}_{k,A}$ is associated current density:

$$\rho_k = |\chi_k|^2 = {A_k}^2 \tag{39}$$

$$\boldsymbol{J}_{k,A} = \frac{{A_k}^2(\nabla_A S_k)}{M_A} \tag{40}$$

This relation does not depend on $\hbar^2$. If we take classical limit in observation, which implies that $\hbar \to 0$, one term from equation (35) equals zero:

$$\frac{\partial S_k}{\partial t} + \sum_{A=1}^{M} \frac{1}{2M_A} (\nabla_A S_k)^2 + E_k = 0 \tag{41}$$

This equation can be written in Hamilton–Jacobi formulation: [14]

$$\frac{\partial S_k}{\partial t} + H_k(\{\boldsymbol{R_A}\}\{\nabla_A S_k\}) = 0 \tag{42}$$

with the classical Hamilton function:

$$H_k(\{\boldsymbol{R_A}\}, \{\boldsymbol{P_A}\}) = T(\{\boldsymbol{P_A}\}) + V_k(\{\boldsymbol{R_A}\}) \tag{43}$$

$\{\boldsymbol{R_A}\}$ represents the set of generalized coordinates, and $\{\boldsymbol{P_A}\}$ their conjugate canonical momenta. The energy for given system is conserved:

$$\frac{dE_k^{tot}}{dt} = 0 \tag{44}$$

$$\frac{\partial S_k}{\partial t} = -(T + E_k) = -E_k^{tot} = const. \tag{45}$$

Following the classically defined Hamilton function (equation (42)), it is concluded that:

$$\boldsymbol{P_A} = \nabla_A S_k = M_A \frac{\boldsymbol{J}_{k,A}}{\rho_k} \tag{46}$$

Newtonian equations of motion with Hamiltoni-Jacobi equation (40) leads us to:

$$\frac{d\boldsymbol{P_A}}{dt} = -\nabla_A E_k \tag{47}$$

$$\frac{d^2\boldsymbol{R_A}(t)}{dt^2} = -\nabla_A V_k^{BO}(\{\boldsymbol{R_A}(t)\}) \tag{48}$$

for each decoupled electronic state *k*. The nuclei move in an effective potential V_k^{BO}, called Born-Oppenheimer potential, which is given from potential energy surface obtained by solving Schrödinger equation for k^{th} state and given nuclear configuration as described above.[14] The potential obtained by calculating the interactions of all particles of the observed system for fixed nuclei positions is used to generate the next nuclei positions. It is independent of the numerical integration step used in simulation. This variant of *ab initio* molecular dynamics is called "Born-Oppenheimer molecular dynamics"[16].

3.2.1. Numerical integration

Because the potential used to generate new nuclei positions in molecular dynamics is not analytically known, generation of new nuclei positions is done by numerical integration. The first step in solving this problem is defining an integration step, by which new nuclei positions will be generated. If the defined step is too small, the duration of calculation will be too long.

In opposite, when the integration step is too large, the credibility of molecular dynamic simulation is compromised. For instance, if the integration step is longer than time of one average vibration, molecular dynamics simulation will be useless for obtaining the properties of studied system. The generated nuclei positions with defined integration time step can be expressed as a set:

$$\{\boldsymbol{R}_A(0)\},\{\boldsymbol{R}_A(0+\Delta t\},\{\boldsymbol{R}_A(2\Delta t)\},\{\boldsymbol{R}_A(3\Delta t)\},\cdots,\{\boldsymbol{R}_A(n\Delta t)\} \tag{49}$$

in which $\boldsymbol{R}_A(0)$ is a set of vectors defined from nuclei positions at the beginning of simulation, Δt is integration step, and *n* is the number of steps for which the calculation will be proceeded. From the equation (49) it is obvious that higher the *n* and Δt it will result with longer computational time for simulation. There are many algorithms for numerical integration which have property of time reversibility. Time reversibility allows us to go back to previous steps after some time of calculating. If we calculate nuclei positions in $t = 0$, $t = \Delta t$, $t = 2\,\Delta t$, $t = 3\,\Delta t$ and so on, it is possible at some step of the simulation to go back to previous steps, for example $t = 2\,\Delta t$. The total time of molecular dynamics simulation for smaller systems is usually around 1 femtosecond (fs). The integration step is between 0,5 fs and 20 fs.

The most algorithms are based on developing the positions of nuclei into Taylor series. The Taylor series of a function is an infinite sum of terms that are expressed in terms of the function's derivatives at a single point[18]. The algorithms that are using Taylor series in numerical integrations do not possess time reversibility.

The most used algorithm with time reversibility is Verlet-Störmer algorithm[19] (VS algorithm). VS algorithm uses positions and accelerations at time *t* and previous positions at time $t - \Delta t$ to generate new positions at time $\Delta t + t$. The positions of nuclei at these moments are expressed as Taylor series at time *t*:

$$\boldsymbol{r}_A(t+\Delta t) = \boldsymbol{r}_A(t) + \Delta t\,\boldsymbol{v}_A(t) + \frac{1}{2}(\Delta t)^2\boldsymbol{a}_A(t) + \cdots + \frac{1}{n!}(\Delta t)^n\frac{d^2\,\boldsymbol{r}_A(t)}{d\,t^n} \tag{50}$$

$$\boldsymbol{r}_A(t-\Delta t) = \boldsymbol{r}_A(t) - \Delta t\,\boldsymbol{v}_A(t) + \frac{1}{2}(\Delta t)^2\boldsymbol{a}_A(t) - \cdots + \frac{1}{n!}(\Delta t)^n\frac{d^2\boldsymbol{r}_A(t)}{d\,t^n} \tag{51}$$

In general, only first three terms are taken into the calculation. Equations (50) and (51) are summed up:

$$\boldsymbol{r}_A(t+\Delta t) = 2\boldsymbol{r}_A(t) - \boldsymbol{r}_A(t-\Delta t) + (\Delta t)^2\boldsymbol{a}_A(t) \tag{52}$$

and the equation for calculating nuclei positions at $t + \Delta t$ from positions at t and $t - \Delta t$ is obtained. The equation (52) does not have velocities of nuclei, which can be calculated from difference in nuclei positions and time interval:

$$\boldsymbol{v_A}(t) = \frac{\boldsymbol{r_A}(t + \Delta t) - \boldsymbol{r_A}(t - \Delta t)}{2\Delta t} \tag{53}$$

For better calculation of the velocities, the equation (50) and (51) should be considered by deriving the function that express position of nuclei in time.

$$\boldsymbol{v_A}(t + \Delta t) = \boldsymbol{v_A}(t) + \Delta t \boldsymbol{a_A}(t) + \frac{1}{2}\boldsymbol{v_A}(\Delta t)^2 \boldsymbol{b_A}(t) + \cdots \tag{54}$$

$$\boldsymbol{v_A}(t - \Delta t) = \boldsymbol{v_A}(t) - \Delta t \boldsymbol{a_A}(t) + \frac{1}{2}\boldsymbol{v_A}(\Delta t)^2 \boldsymbol{b_A}(t) - \cdots \tag{55}$$

By repeating the same procedure explained for getting $\boldsymbol{r_A}(t + \Delta t)$, after summing up equations (54) and (55), velocities are:

$$\boldsymbol{v_A}(t + \Delta t) = 2\,\boldsymbol{v_A}(t) - \boldsymbol{v_A}(t - \Delta t) + (\Delta t)^2 \boldsymbol{b_A}(t) \tag{56}$$

The third term in equation (56) is derivative of acceleration and sometimes it can be excluded. The velocities are important factor in monitoring the fluctuations in total kinetic energy through simulation. [20] Usually, the initial velocities for particles at $t = 0$ are manually defined for each particle, or randomly based on the temperature of simulation. If the temperature is used for assigning initial velocities, Maxwell Boltzmann distribution of velocity components is used:

$$\rho(v_{x,i}) = \frac{1}{\sqrt{2\pi\sigma^2{}_{x,i}}} e^{\frac{-v_{x,i}^2}{2\sigma^2{}_{v,i}}} \tag{57}$$

$$\rho(v_{y,i}) = \frac{1}{\sqrt{2\pi\sigma^2{}_{y,i}}} e^{\frac{-v_{y,i}^2}{2\sigma^2{}_{v,i}}} \tag{58}$$

$$\rho(v_{z,i}) = \frac{1}{\sqrt{2\pi\sigma^2{}_{z,i}}} e^{\frac{-v_{z,i}^2}{2\sigma^2{}_{v,i}}} \tag{59}$$

in which T represents thermodynamic temperature at which simulation is performed, m_i the mass of the individual nucleus, and k_b Boltzmann constant.

Variance $\sigma_{v,i}^2$ is given as:

$$\sigma_{v,i}^2 = \frac{k_b T}{m_i} \tag{60}$$

In addition to determining initial velocities, the total angular momentum is set to zero in order to avoid system translation in space during the simulation. The stability of simulation is hard to follow since the error is induced to calculation from velocities obtained in this way. Molecular

dynamics simulation stability can be followed by calculating Root Mean Square factor (RMS factor) for energy. If there are fluctuations in RMS factor, simulation is not stable, so properties or/and procedures used in simulation should be changed.

The variant of the Verlet-Störmer algorithm that gives better results and ability to follow the simulation stability is Beeman's algorithm. Beeman's algorithm in prediction of nuclei positions is using the term with velocities:

$$\boldsymbol{r_A}(t+\Delta t) = \boldsymbol{r_A}(t) + \boldsymbol{v_A}(t)\Delta t + \frac{1}{6}\big(4\boldsymbol{a_A}(t) - \boldsymbol{a_A}(t-\Delta t)\big)(\Delta t)^2 \tag{61}$$

$$\boldsymbol{v_A}(t+\Delta t) = \boldsymbol{v_A}(t) + \frac{1}{6}\big(2\boldsymbol{a_A}(t+\Delta t) + 5\boldsymbol{a_A}(t) - \boldsymbol{a_A}(t-\Delta t)\big)\Delta t \tag{62}$$

The disadvantage of Beeman's algorithm is increasing the time of computation.

3.3. Machine Learning

Machine learning[21] (**ML**) is a field of computer science that uses statistical techniques to give computers the ability to learn with data, without being explicitly programmed. ML has many sub fields. Some sub fields are statistical learning methods, neural networks, computational learning theory, and data mining.[22] The advantage of ML algorithms is the capability of machine or software to improve its performance through experience. A typical ML model learns the knowledge from data it is exposed to and then applies it to new problems. There are three types of machine learning: supervised ML, unsupervised ML, and reinforcement learning.

Supervised ML[23] is when computer learns a model from labelled data (called training data). Training data allow machine to make predictions about new unseen data. It is called supervised ML because for the labelled training data outputs are already known. The model learning can be carried out using discrete class or continuous class labels. If using discrete class, it is called classification and for continuous regression ML.

Unsupervised ML[23] is the opposite of supervised ML. The input for unsupervised ML algorithms is unlabelled data or data of unknown structure. Output data are not provided so these types of algorithms are trying to find some regularities in input data. Often data contain patterns which lead to extracting the meaningful data as output. Most widely used technique is clustering, which allows grouping data into piles that have meaningful information.

Reinforcement learning[24] is used in systems where output is a sequence of actions. For reaching the goal, sequence of correct actions is important, since one single action means nothing. Reinforcement algorithm learns from past actions that led to caused goal. Sometimes,

reinforcement learning can be considered as supervised ML. There are some examples of possible applications in the following tables.

Table 1. Types of ML and their application.[25]

supervised ML	unsupervised ML	reinforcement ML
classification o fraud detection o e-mail spam detection o diagnostics *image classification* o regression o risk assessment o score prediction	*dimensionality reduction* o text mining o face recognition o big data o image visualization *image visualization* o clustering o biology and chemistry o city planning o targeted marketing	o gaming o finance sector o manufacturing o inventory management o robot navigation

3.3.1. Linear (In)dependence of Vectors

A set of vectors is linearly dependent if there is at least one vector from that set which can be written as a linear combination of others. If there is no vector that can be written as a linear combination of the other vectors, the set is linearly independent. Two vectors are linearly independent if they are satisfying equation[26]:

$$a\boldsymbol{u} + b\boldsymbol{v} = 0 \tag{63}$$

in which $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors, a and b being constants. Usually, vectors are written in matrix representation as:

$$\boldsymbol{u} = \begin{bmatrix} x \\ y \end{bmatrix} \tag{64}$$

$$\boldsymbol{v} = \begin{bmatrix} z \\ q \end{bmatrix} \tag{65}$$

After putting equations (63) and (64) into equation (62):

$$\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} z \\ q \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{66}$$

the homogenous system of equations is given. There are multiple methods that can be used for solving this type of equations. The most used method is Gaussian elimination, which implies the reduction of Gauss matrix. The Gauss matrix can be written as:

$$\begin{bmatrix} x & z & 0 \\ y & q & 0 \end{bmatrix} \tag{67}$$

When matrix is reduced, the set of vectors is linearly independent if the matrix does not contain null vectors.

The other possible method to find out if the set of vectors is independent is to find inverse matrix of initial matrix of vectors.[27] Between the initial matrix and its inverse matrix, the following equation is valid:

$$\boldsymbol{M}^{-1}\boldsymbol{M} = \boldsymbol{M}\boldsymbol{M}^{-1} = \boldsymbol{I} \tag{68}$$

in which $\boldsymbol{M}$ represent the initial matrix, $\boldsymbol{M}^{-1}$ its inverse, and $\boldsymbol{I}$ identity matrix. For the identity matrix, diagonal elements are equal to 1, and non-diagonal to 0. If the determinant of matrix equals zero, it is not possible to find its inverse matrix, so the set of vectors is linearly dependent. Also, for calculating the inverse matrix, initial matrix should be square $n \times n$. Linear dependence of vectors is explained on a 2-dimensional problem, but all mentioned equations are valid for vectors with higher dimensionality. The main problem of this method is that it requires the square matrix.

3.3.2. Eigendecomposition of the Matrix

One of the most widely used matrix decomposition is called eigendecomposition[27], in which a matrix is decomposed into a set of eigenvectors and eigenvalues.[28] If we have a square matrix $\boldsymbol{A}$, eigenvector of matrix $\boldsymbol{A}$ is nonzero vector $\boldsymbol{v}$:

$$\boldsymbol{A}\,\boldsymbol{v} = \lambda\,\boldsymbol{v} \tag{69}$$

The scalar λ is called eigenvalue and it is corresponding to eigenvector $\boldsymbol{v}$. Any rescaled vector $s\boldsymbol{v}$ has the same eigenvalue as $\boldsymbol{v}$. Usually, unit eigenvectors are looked for. If we suppose that matrix $\boldsymbol{A}$ has n linearly independent eigenvectors with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, we can concatenate all the eigenvalues into a new matrix $\boldsymbol{V}$ with one eigenvector per column:

$$\boldsymbol{V} = \left[\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots, \boldsymbol{v}^{(n)}\right] \tag{70}$$

Also, we can concatenate eigenvalues in same way:

$$\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_n]^T \tag{71}$$

The eigendecomposition of $\boldsymbol{A}$ is then:

$$\boldsymbol{A} = \boldsymbol{V}\,\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{V}^{-1} \tag{72}$$

Constructing matrices with eigenvectors and their eigenvalues enables us to stretch space in desired direction. It is not possible to decompose every matrix into eigenvectors and eigenvalues, sometimes the decomposition exists but it is harder to be obtained because it

involves complex rather than real numbers. Every symmetric matrix can be decomposed using only real valued eigenvectors and eigenvalues:

$$\boldsymbol{A} = \boldsymbol{Q}\,\boldsymbol{\Lambda}\,\boldsymbol{Q}^T \tag{73}$$

where $\boldsymbol{Q}$ is orthogonal matrix composed of eigenvectors od $\boldsymbol{A}$, $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues.

The eigendecomposition for symmetric matrices may not be unique. If two or more eigenvectors have the same eigenvalue, then any set of orthogonal vectors lying in their span are eigenvectors with that values.[28] By convention, entries of eigenvalues are sorted in descending order and the eigendecomposition is unique only if all the eigenvalues are unique.

3.3.3. Singular Value Decomposition

Singular Value Decomposition (SVD)[28] is another way of factorizing a matrix into singular vectors and singular values. Every real matrix has an SVD, which is not true for eigendecomposition. If the matrix is not a square matrix, the eigendecomposition is not defined, but SVD can still be calculated. Procedure is like eigendecomposition (equation (71)), but instead of diag (λ), matrix is used:

$$\boldsymbol{A} = \boldsymbol{U}\,\boldsymbol{D}\,\boldsymbol{V}^T \tag{74}$$

Matrix $\boldsymbol{A}$ is an $m \times n$ matrix, $\boldsymbol{U}$ is $m \times m$, $\boldsymbol{D}$ is $m \times n$ and $\boldsymbol{V}$ is $n \times n$ matrix. The matrices $\boldsymbol{V}$ and $\boldsymbol{U}$ are defined to be orthogonal matrices, $\boldsymbol{D}$ is defined to be a diagonal matrix. The diagonal elements of matrix $\boldsymbol{D}$ are known as the singular values of matrix $\boldsymbol{A}$. The columns or $\boldsymbol{V}$ are called right-singular vectors. The SVD of $\boldsymbol{A}$ can be interpreted in the terms of the eigendecompositions of functions of $\boldsymbol{A}$.

3.3.4. The Moore-Penrose Pseudoinverse

Matrix inversion is not defined for matrices that are not square. The following equation should be solved:

$$\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{y} \tag{75}$$

If dimension of matrix $\boldsymbol{A}$ is $m \times n$ and $m>n$ relation is valid, pseudoinverse of matrix[28] $\boldsymbol{A}$ is defined as:

$$\boldsymbol{A}^{+} = \lim_{\alpha \to 0} (\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A} + \alpha \boldsymbol{I})^{-1}\,\boldsymbol{A}^T \tag{76}$$

Algorithms for calculating pseudoinverse are usually based on equation:

$$\boldsymbol{A}^{+} = \boldsymbol{V}\,\boldsymbol{D}^{+}\,\boldsymbol{U}^T \tag{77}$$

$\boldsymbol{V}$, $\boldsymbol{D}$ and $\boldsymbol{U}$ are the SVD of $\boldsymbol{A}$, and the pseudoinverse $\boldsymbol{D}^{+}$ of a diagonal matrix $\boldsymbol{D}$ is obtained by taking the reciprocal of its nonzero elements then transposing of the resulting matrix. For $n > m$, solving the equation provides one of the many possible solutions.

3.3.5. Principal Component Analysis

Principal Component Analysis (PCA) is a tensor decomposition method, often used in data reduction, classification and grouping of observations and modelling relationships that may exist between variables.[29] It is a descriptive method that provides geometric representation. Also, it is the most used method in a process of data mining since it is quite simple and non-parametric method.[29] Data mining implies transformation of raw information to useful information. Usually, the data obtained in scientific experiments are clouded, redundant and unclear. That makes people unable to see the connections between individual variables. Methods that can reduce dimensionality and/or group data are needed to solve that problem. They can extract something hidden in data that can lead us to a better and correct conclusion.

The first assumption in PCA is linearity.[30] It limits the re-expressing data as a linear combination of its basis vectors. Every next vector of a basis set needs to be perpendicular to the previous one.

First, we assume that our original data is written as matrix $\boldsymbol{X}$ and the goal of the PCA procedure is to change the basis set of $\boldsymbol{X}$ as:

$$\boldsymbol{PX} = \boldsymbol{Y} \tag{78}$$

where $\boldsymbol{P}$ is matrix that transforms the $\boldsymbol{X}$ into matrix $\boldsymbol{Y}$. Matrix $\boldsymbol{P}$ is matrix that is geometrically stretching and rotating matrix $\boldsymbol{X}$ into matrix $\boldsymbol{Y}$. The rows of $\boldsymbol{P}$ are new basis vectors, the principal components of $\boldsymbol{X}$. That can be expressed as:

$$\begin{bmatrix} \boldsymbol{p_1} \\ \vdots \\ \boldsymbol{p_m} \end{bmatrix} [\boldsymbol{x_1} \quad \cdots \quad \boldsymbol{x_n}] = \begin{bmatrix} \boldsymbol{p_1} \cdot \boldsymbol{x_1} & \cdots & \boldsymbol{p_1} \cdot \boldsymbol{x_n} \\ \vdots & \ddots & \vdots \\ \boldsymbol{p_m} \cdot \boldsymbol{x_1} & \cdots & \boldsymbol{p_m} \cdot \boldsymbol{x_n} \end{bmatrix} \tag{79}$$

Each column of $\boldsymbol{Y}$ is $\boldsymbol{y_i}$:

$$\boldsymbol{y_i} = \begin{bmatrix} \boldsymbol{p_1} \cdot \boldsymbol{x_i} \\ \vdots \\ \boldsymbol{p_m} \cdot \boldsymbol{x_i} \end{bmatrix} \tag{80}$$

The j^{th} coefficient of $\boldsymbol{y_i}$ represents projection on the j^{th} row of $\boldsymbol{P}$.

To obtain the best choice of a basis set, the noise in measurement data needs to be low in any basis set. The noise minimization can be done by maximizing the variance. We can define signal-to-noise ratio as a ratio of variance of signal and variance of noise. Higher signal-to-

noise ratio means high precision measurement. Therefore, the noise minimization can be done by maximizing the variance of the signal.[30]

The other problem that appears is redundancy. This problem can be solved by calculating the covariance matrix. Covariance matrix can be expressed as:

$$\boldsymbol{C_X} = \frac{1}{n}\boldsymbol{XX^T} \tag{81}$$

where n is the number of samples. Each column of X corresponds to a set of measurements from one particular trial. $\boldsymbol{C_X}$ is a sqaure matrix ($m \times m$), in which ij^{th} element is the dot product between the vector of the i^{th} measurement type with the vector of the j^{th} measurement type. The next step is decomposing the covariance matrix into its eigenvectors and eigenvalues.

$$\boldsymbol{C_Y} = \boldsymbol{PC_XP^T} \tag{82}$$

In a manipulated covariance matrix, all non-diagonal elements should be equal to zero, so the $\boldsymbol{Y}$ is decorrelated. The easiest way to do that is to assume that, while decomposing, all basis vectors are orthonormal. The first vector of a basis set is chosen in a m-dimensional space, along which the variance in $\boldsymbol{X}$ is maximized. Next vectors of a basis set are chosen to represent the maximum of variance in $\boldsymbol{X}$ but with orthonormality condition to all previously chosen vectors of a basis set. The procedure is repeated until most of the variance of $\boldsymbol{X}$ is described by basis set vectors. Usually, 90% of the variance is enough to describe the initial data without important losses. The vectors of final basis sets are called *Principal Components* (PCs).

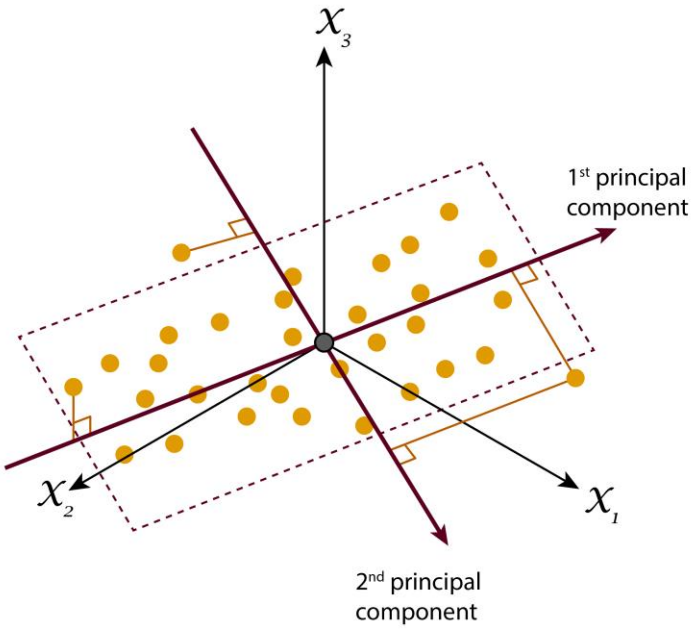


Figure 6. An example of PCA analysis result.

Figure 6. shows the result of PCA for the data that was depending on three variables (χ_1, χ_2 and χ_3). The PCA analysis reduced the dimension so the data is described by two principal components. First one that describes the biggest part in variance is called the first PC, next one the second PC and so on. A graphical representation of data in principal components basis set can be a visual indication for grouping data according to some properties. This type of graphical representation is called *score-plot*.

3.3.6. QR Decomposition

The QR decomposition of factorization of an $n \times m$ matrix $\boldsymbol{A}$ assumes the following form:

$$\boldsymbol{A} = \boldsymbol{QR} \tag{83}$$

where **Q** is an $n \times n$ orthogonal matrix, and $\boldsymbol{R}$ is:

$$\boldsymbol{R} = \boldsymbol{Q^T A} \tag{84}$$

and has zeros on elements below its diagonal. If n is equal or greater than m, then it can be written as:

$$\boldsymbol{Q^T A} = \begin{bmatrix} \boldsymbol{R_{11}} \\ 0 \end{bmatrix} \tag{85}$$

where $\boldsymbol{R_{11}}$ is an $n \times n$ upper triangular matrix.

If it contains linearly independent columns, it can be factored as:

$$\boldsymbol{A} = [\boldsymbol{q_1} \quad \cdots \quad \boldsymbol{q_n}] \begin{bmatrix} \boldsymbol{R_{11}} & \cdots & \boldsymbol{R_{1n}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{R_{nn}} \end{bmatrix} \tag{86}$$

In opposite, for m greater than n, QR decomposition of $\boldsymbol{A}$ is:

$$\boldsymbol{Q^T A} = [\boldsymbol{R_{11}} \quad \boldsymbol{R_{12}}] \tag{87}$$

In equation (87) all diagonal elements $\boldsymbol{R_{ii}}$ are non-zero and most definitions require $\boldsymbol{R_{ii}}$ greater than zero. That makes $\boldsymbol{Q}$ and $\boldsymbol{R}$ unique. $\boldsymbol{Q}$ is $m \times n$ with orthonormal columns. If the $\boldsymbol{A}$ is square matrix than $\boldsymbol{Q}$ is orthogonal and equal to $\boldsymbol{I}$ matrix. $\boldsymbol{Q}$ and $\boldsymbol{R}$ are called *Q-* and *R-factors*. The most known methods to calculate $\boldsymbol{Q}$ and $\boldsymbol{R}$ are Gram–Schmidt process, Householder transformations, or Givens rotations.

In case of using Gram-Schmidt process, first the matrix $\boldsymbol{Q}$ is calculated. The matrix $\boldsymbol{Q}$ is orthonormal, by rules of Gram-Schmidt process. The matrix $\boldsymbol{R}$ is then calculated from matrices $\boldsymbol{Q}$ and $\boldsymbol{A}$ by following equation (84). The Householder transformations are transformations that take a vector and reflect it about some (hyper)plane. Matrix $\boldsymbol{Q}$ is used in a way that all coordinates, but one, disappear. It is used multiple times until upper triangular form is obtained.

Givens rotations are used to zero elements in the subdiagonal of the matrix, forming the $\boldsymbol{R}$ matrix. The orthogonal $\boldsymbol{Q}$ matrix is concatenation of all the Givens rotations. All mentioned methods have their own advantages and disadvantages.

§ 4. EXPERIMENTAL SECTION

4.1. *Ab initio* Molecular Dynamics Simulation

Molecular dynamics run of 5 000 000 steps was simulated using *on-the-fly* calculations of forces. The forces were calculated using the PM7 method implemented in MOPAC2016.[7] To ensure adequate sampling of the phase space, initial velocities were selected from Maxwell distribution at 1273.15 K. The temperature was kept constant during the simulations using a velocity scaling algorithm. The step size was 0.5 fs and total length of simulation was 2.5 ps. This total simulation length was sufficient for full conformational analysis, which is confirmed by the calculation of strict local maxima *plateaus*.

These points represented a sampling space from which the initial guess structures for conformational analysis were extracted. This extraction was performed by finding all strict local maxima in the probability distribution of the molecular geometry coordinates. To be more precise, it was reasonable to expect that, during the molecular dynamics run, the molecule would statistically spend more time in and around the minima points on the potential energy surface and that, consequently, the probability distribution for the molecular structures in these areas of the phase space would have a strict local maximum. In fact, the search for strict local minima on a potential energy surface is equivalent to the search for strict local maxima in a probability distribution of molecular geometry coordinates. And these coordinates can be defined in any possible way.

Because the dimensionality of this search for strict local minima problem is of the order $3N$ (or $3N–6$ if the search is performed in the space of internal coordinates), where N is the number of atoms in the molecule, it was reasonable to make a reduction of the problem to fewer dimensions. All molecular dynamics simulations were performed using the quantum chemistry code ***qcc***.[8]

4.2. Machine Learning Determination of Internal Coordinate Distances

The trajectory of (*R*)-cinchonidine in the Cartesian coordinates was transformed to trajectory in all possible internal coordinate distances. The (*R*)-cinchonidine has 44 atoms, and the number of all possible generalized internal coordinate distances between all atoms is 946. This corresponds to the number of all atom pairs $\binom{44}{2}$ that can be selected in a molecule. The file

with the definition of all 946 internal coordinate distances was also generated. The transformation was carried out using the ***moonee*** program code.[9]

The progressive machine learning algorithm was applied on a given trajectory using the different number of points. Number of points was gradually increased, starting from 1000 to 10 000 points (with a chunk of 1000 points), and from 10 000 to 100 000 points (with a chunk of 10 000 points) to determine the total number of generalized coordinates and the real time of calculation. These generalized coordinates will be linearly independent and appropriate for defining a molecule's geometry.

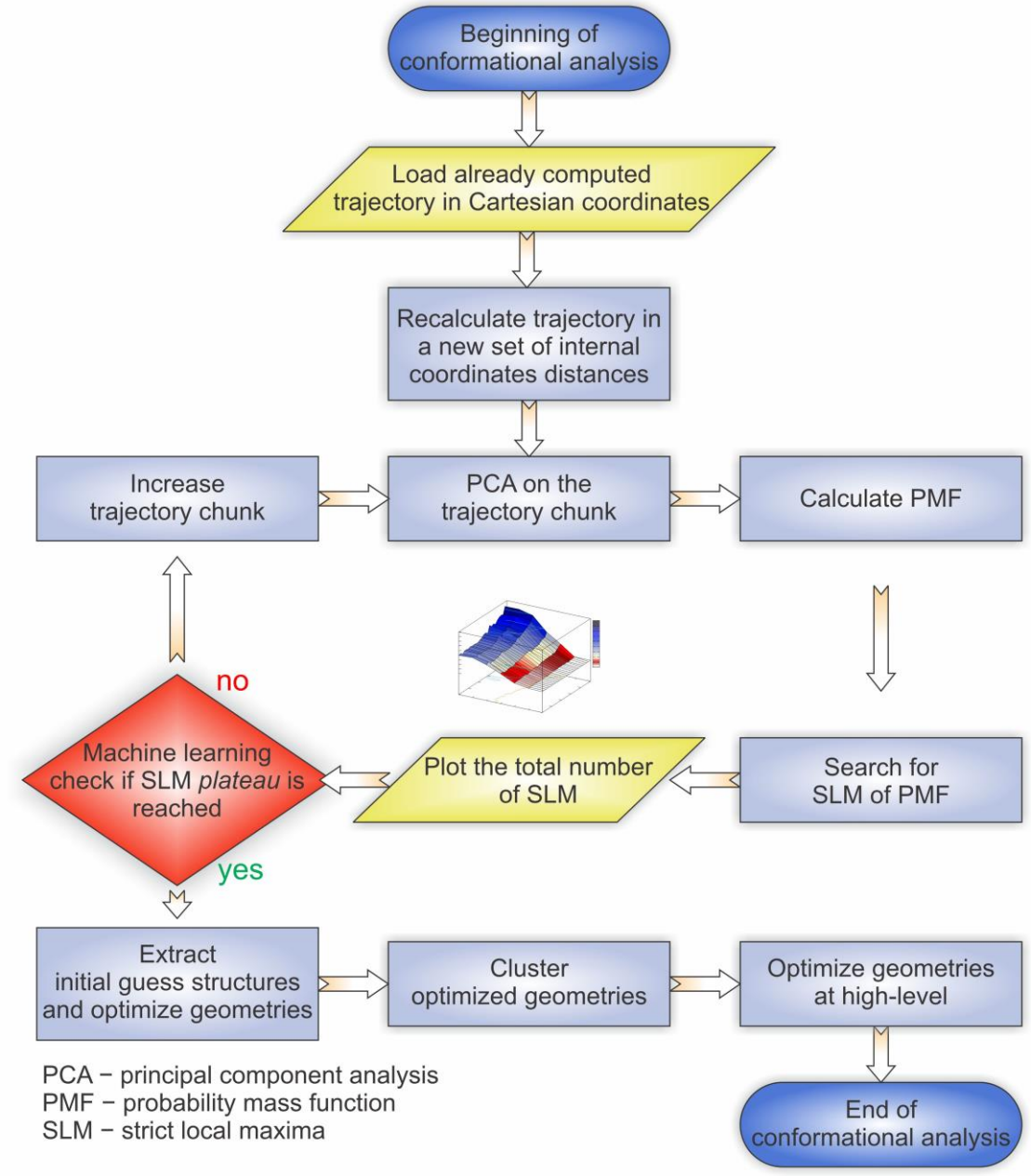
Different sets of molecular geometries in internal coordinate distances representation were written in matrices and ranks of these matrices were determined. Determination of ranks was performed by QR decomposition where the lead factors in decomposed matrices were taken into account for rank determination. By using *leave-one-row-out* method through the distance dimension of matrices, all significant rows for the description of molecular geometries were determined. If the row with the values obtained from a specific distance definition does not contribute to the overall rank, iterative application of machine learning was utilized in such a way that this row was deleted from the following calculation. If this row contributes to the rank, it was kept in the matrix and in the following calculation.

In this way, the most optimal representations of distances for different lengths of simulation were determined and then tested for convergence. As previously described, lengths of simulation were selected starting from 1000 and continued up to 10 000 by a 1000-point step. Investigation of intermediate gradients provided information about overall increase of the number of defined distances, and at that point the chunk was increased to 10 000 and scanned up to the 100 000 points of simulation, where convergence was obtained.

4.3. Calculation of Strict Local Maxima *Plateaus*

To confirm the adequacy of the present set of generalized internal coordinate distances, strict local maxima *plateaus* were used. SLM *plateaus* were calculated by the well-established procedure described in previously published work.[6,13,31] After loading up of already calculated trajectory described in Cartesian coordinates, trajectory was recalculated using the newly defined internal coordinates distances determined with the procedure described in 4.2. An iterative machine learning procedure consisting of principal component analysis and the search for strict local maxima was utilized to identify the area of SLM-*plateau*. Initially, PCA was calculated for one chunk of trajectory (250 000 points) and SLM were plotted. Length of the

trajectory was increased for an additional chunk of 250 000 points and the procedure was repeated until the convergence was achieved. The workflow diagram of the procedure is presented in Scheme 1.



Scheme 1. Calculation of strict local maxima *plateau* using the tensor decomposition of molecular dynamics trajectories.

4.4. Computational Resources

All calculations were conducted on a low-cost workstation equipped with Intel(R) Core(TM)2 Duo processor with two cores operating at 3.00 GHz. Since most of the coded procedures is parallelized, both cores were used where possible. Total memory of the workstation was 8 Gb, which was not sufficient in some cases causing the paging and consequently – slowing down the calculation. This was particularly noticeable with the bigger sizes of trajectory. For this reason, the real time of computation was not available for the trajectory sizes bigger that 30 000 points.

§ 5. RESULTS AND DISCUSSION

5.1. *Ab initio* Molecular Dynamics Simulation

Ab initio molecular dynamics simulation of (*R*)-cinchonidine was previously conducted and a total of 5 000 000 steps was simulated.[7] Forces on all atoms were calculated *on-the-fly* using the PM7 method implemented in MOPAC2016. Since the final results given in Ref. 7 present the full conformational space of (*R*)-cinchonidine, phase space was adequately covered. This was achieved by sampling of the phase space using the velocity scaling in molecular dynamics simulation that kept the temperature of 1273.15 K during simulation. Step size was 0.5 fs ensuring smooth transitions in the potential energy surface. Total length of simulation was 2.5 ps, which was excessive according to the SLM-*plateau* criteria, but this guaranteed sufficient sampling for determination of generalized linearly independent set of internal coordinate distances.

5.2. Generalized Internal Coordinate Distances

Extensive machine learning calculation was applied on the determination of linearly independent set of internal coordinate distances. Exact representation in Cartesian coordinates for (*R*)-cinchonidine molecule was already established in our previous work,[6] but determination of adequate and the most optimal set of distances for description of molecular structure is still an open problem. Determination of internal coordinate distances from the full MD trajectory is one possibility, although not quite feasible due to the large number of trajectory points (5 000 000). For the molecule in question, this task is achievable, but for larger molecules ($N > 100$), memory demands are too high. It was reasonable to expect that there must be a point in the simulation after which the set of internal coordinate distances remains more or less constant – providing satisfactory description of the molecular geometry.

The strategy for determination of linearly independent set of internal coordinate distances was the following:

1. determine the set of internal coordinate distances in dependence on the simulation length, and

2. check each set using the tensor decomposition and comparison to already known exact results obtained in Cartesian coordinates[6]

The set of distances was firstly determined for the 1000 points, then 2000 *etc*. The chunk size was 1000 points until the limit of 10 000 points of simulation was reached. Then the chunk size was increased to 10 000 and procedure was continued until the convergence in the number and the definition of internal coordinate distances was reached.

For each investigated length of simulation, the number and the definition of internal coordinate distances was determined and saved. Obtained results are summarized in Table 2, whereas definition of all internal coordinate distances is presented in Tables A1–A19.

Table 2. Total number of linearly independent internal coordinate distances after applying machine learning algorithm on various (*R*)-cinchonidine trajectory lengths.

Label	*N*(points)	*N*(internal coordinate distances)	Numerical gradient of *N*(internal coordinate distances)	*t* / min
a)	1000	43	0.0430	5.2
b)	2000	58	0.0150	14.5
c)	3000	68	0.0100	26.9
d)	4000	79	0.0110	37.4
e)	5000	82	0.0030	47.3
f)	6000	87	0.0050	58.6
g)	7000	93	0.0060	69.0
h)	8000	99	0.0060	93.9
i)	9000	112	0.0130	86.5
j)	10 000	118	0.0060	96.6
k)	20 000	144	0.0026	241.2
l)	30 000	155	0.0011	377.0
m)	40 000	166	0.0011	-
n)	50 000	170	0.0004	-
o)	60 000	201	0.0031	-
p)	70 000	211	0.0010	-
q)	80 000	219	0.0008	-
r)	90 000	220	0.0001	-
s)	100 000	220	0.0000	-

For a very short length of simulation of 1000 points, a total of 43 linearly independent generalized internal coordinate distances was obtained (Table 2). These 43 distances represent the best possible linearly independent description of (*R*)-cinchonidine molecular geometry expressed in the set of internal coordinate distances for this simulation length (Table A1). Increase in the length of simulation to 2000 points resulted in a total of 58 determined internal coordinate distances. Since the increase in the number of coordinates was significant, the size of the investigated trajectory was further increased by the chunk of 1000 points.

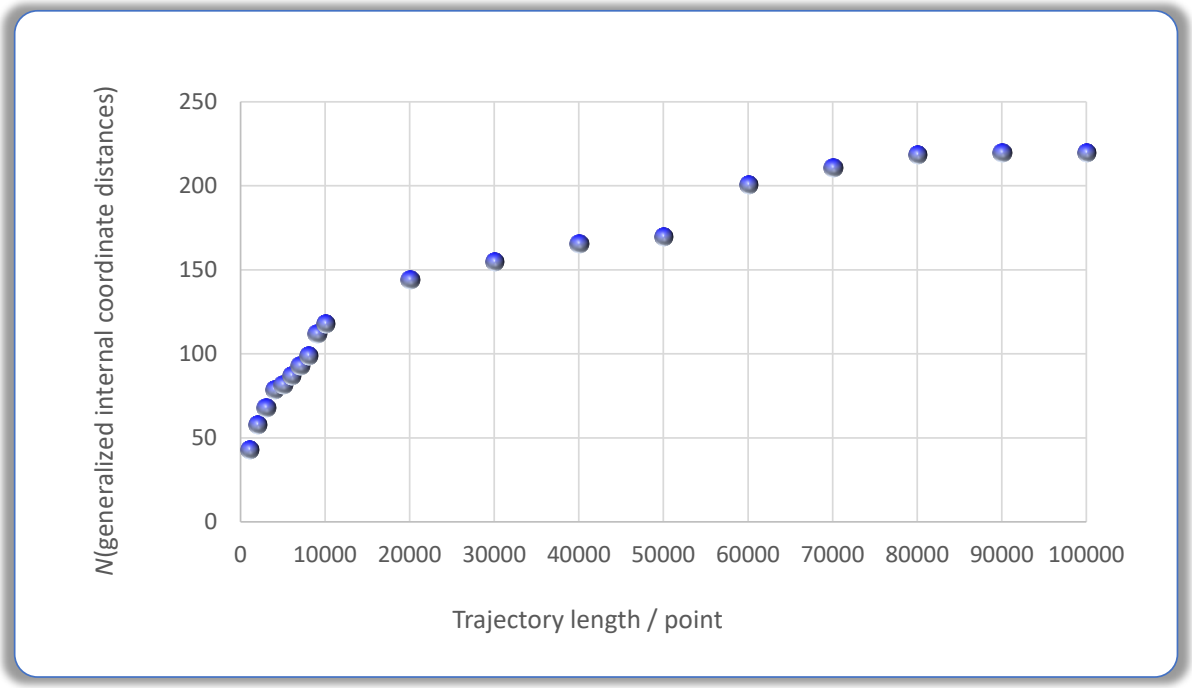


Figure 7. Total number of linearly independent internal coordinates determined for (*R*)-cinchonidine.

Monitoring the progress in the size of the set of generalized internal coordinate distances was evaluated by calculating numerical gradients of the total number of coordinates. Approximately, when the size of simulation trajectory reached 10 000 points, gradients were low enough indicating that the chunk size should be increased (Table 2 and Figure 8).

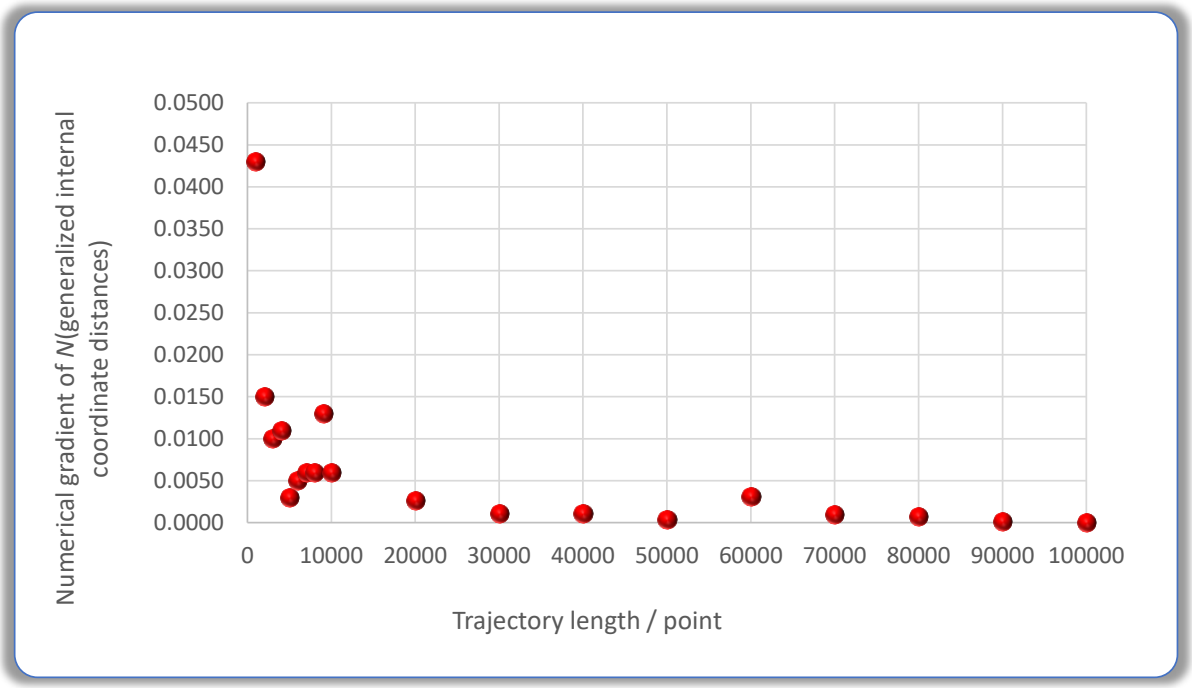


Figure 8. Numerical gradients of the total number of linearly independent internal coordinates determined for (*R*)-cinchonidine.

After point 80 000, the total number of internal coordinate distances converged. Gradients were also very low, indicating the convergence in the number of the internal coordinate distances (Figs. 7 and 8).

Analyzing the progress of machine learning procedure for removing the linear dependency among all defined internal coordinate distances in the molecular dynamics trajectory, an interesting fact can be observed. In the set of all possible 946 defined distances, various distances present linear combinations of some other distances. Selection of the linearly independent set of distances is biased by the definition order. One can see that from the Fig. 9 where the progress of the removal process is presented for two cases. The first one (Fig. 9a) is for the initial size of the trajectory (1000 points), and the second one is for the size of the trajectory where the set of internal coordinate distances was converged (80 000 points, Fig. 9b). Majority of the distances firstly defined are removed from the set due to the linear dependency. But the same could be expected if the order of the internal coordinate distances is reversed (because the firstly defined coordinates in this new order are now linearly dependent on the others, later defined coordinates).

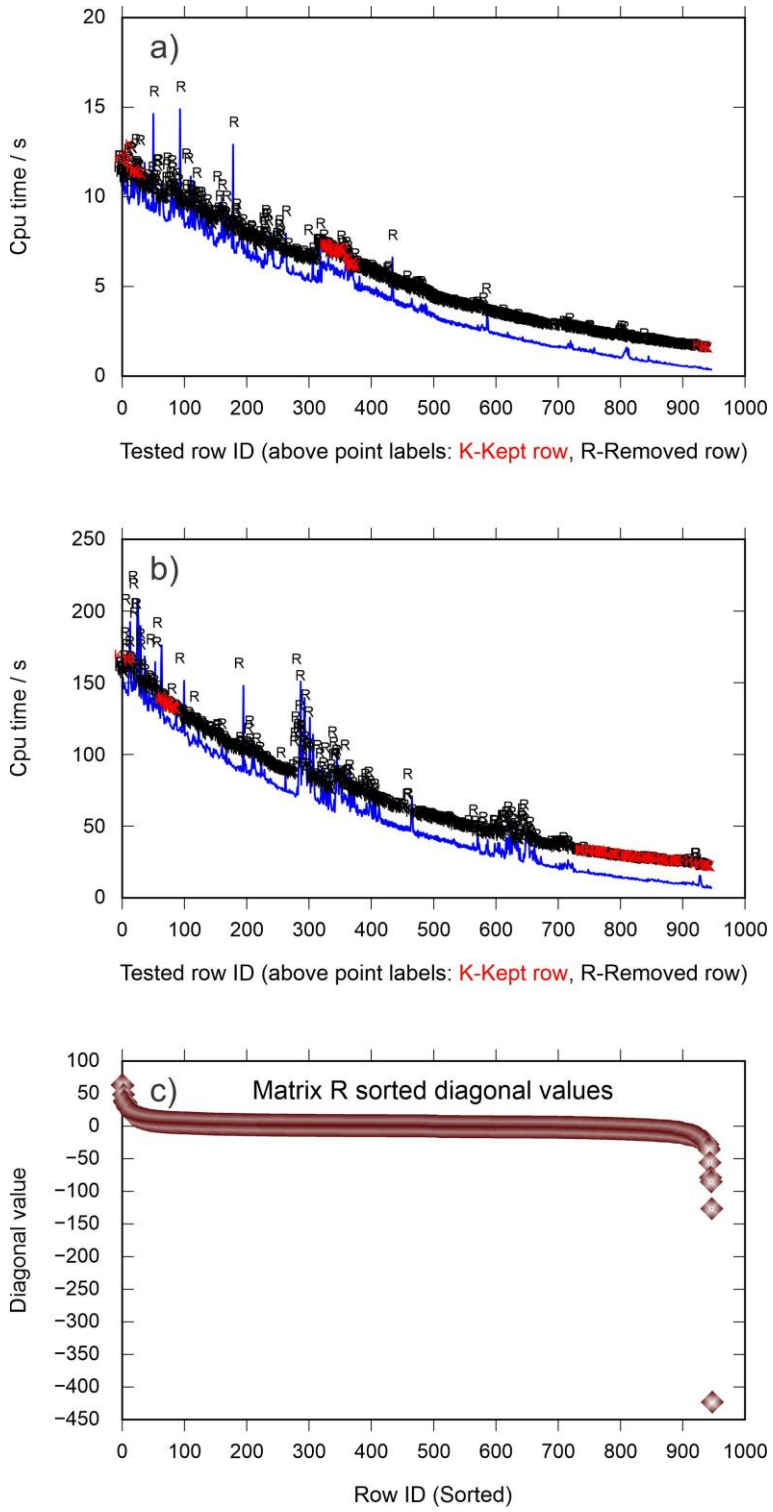


Figure 9. Machine learning progress of the linear dependency removal of internal coordinate distances in the molecular dynamics trajectory (a) 1000 points, b) 80 000 points), and sorted diagonal values of matrix *R*.

On Fig. 9c, the sorted diagonal values, for the size of the trajectory where the set of internal coordinate distances was converged (80 000 points), are presented. The complexity of the rank

determination criteria can be observed from this figure. Majority of the diagonal entries have absolute value very near the zero, making it very difficult to determine the proper rank of the matrix.

The graphical representation of real time of computation on a number of different trajectory lengths is shown on Fig. 10. Linear dependence on the total number of points was expected, and it was obtained for points up to the 30 000. After that point, the lack of server memory caused disk swapping during the QR decomposition that influenced the computational time and these values were not representative. This did not stop the calculation, just slowed it down, but the obtained timings were unrealistic.

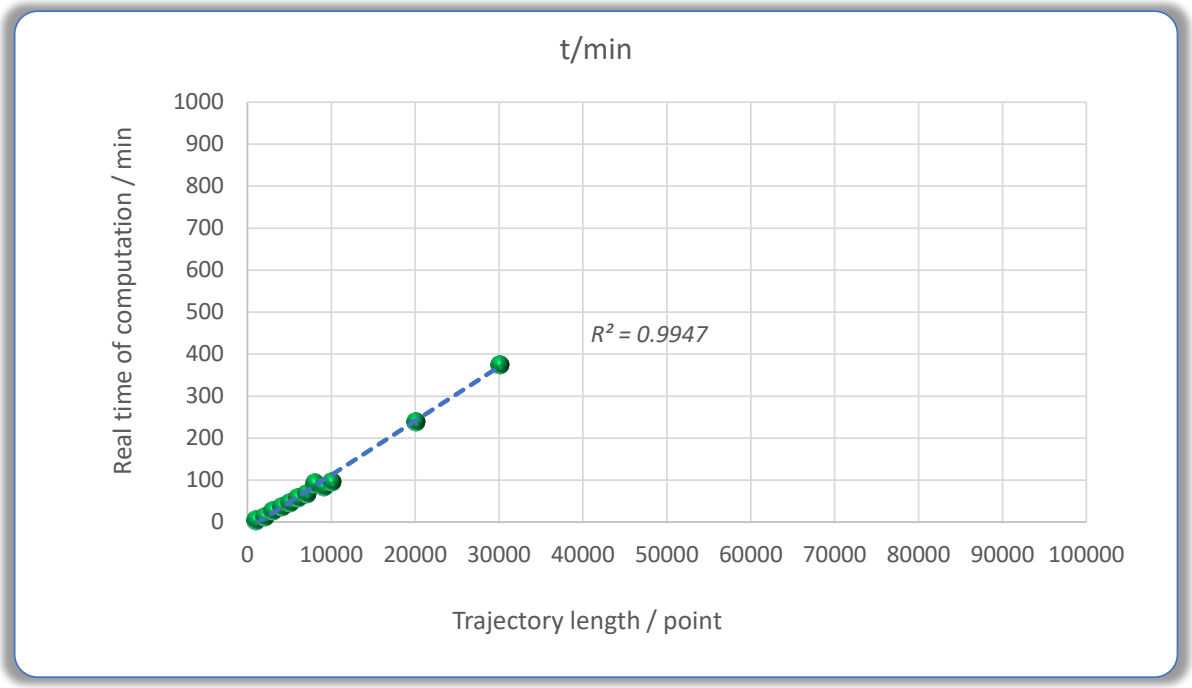


Figure 10. Real time of computation for determination of linearly independent internal coordinate distances.

5.3. Strict Local Maxima *Plateaus*

To confirm that obtained converged set of distances is indeed a completely linearly independent set of internal coordinates suitable for complete representation of the molecular structure, for each determined set of distances, *plateau* of strict local maxima in dependence of number of simulation steps and dimension of reduced space was calculated (Fig. 11.).[13]

For 43 distances determined from the first 1000 points in the simulation, *plateau* was not obtained (Fig. 10a), which was a clear evidence that this set was not appropriate for the description of (*R*)-cinchonidine molecular structure. As expected, for the next investigated size of the trajectory (2000 points), SLM-*plateau* was still not noticeable (Fig. 10b). The same results were obtained for the cases up to 80 000 points (Figs. 10c–10p), where the SLM-*plateau* can be observed. This also corresponds to the convergence in the total number of determined linearly independent internal coordinate distances (Fig. 7 and 8).

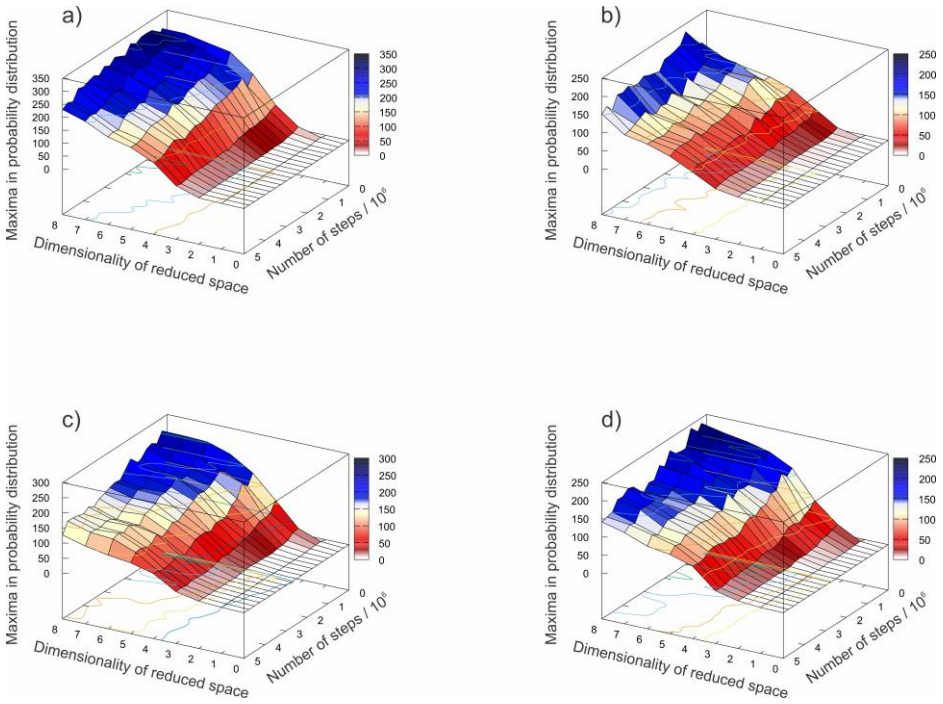


Figure 11. *Plateaus* of strict local maxima in dependence on the total number of linearly independent internal coordinate distances (labels *a*–*s* follow classification given in Table 2, label *t* corresponds to the exact representation and is given for comparison).

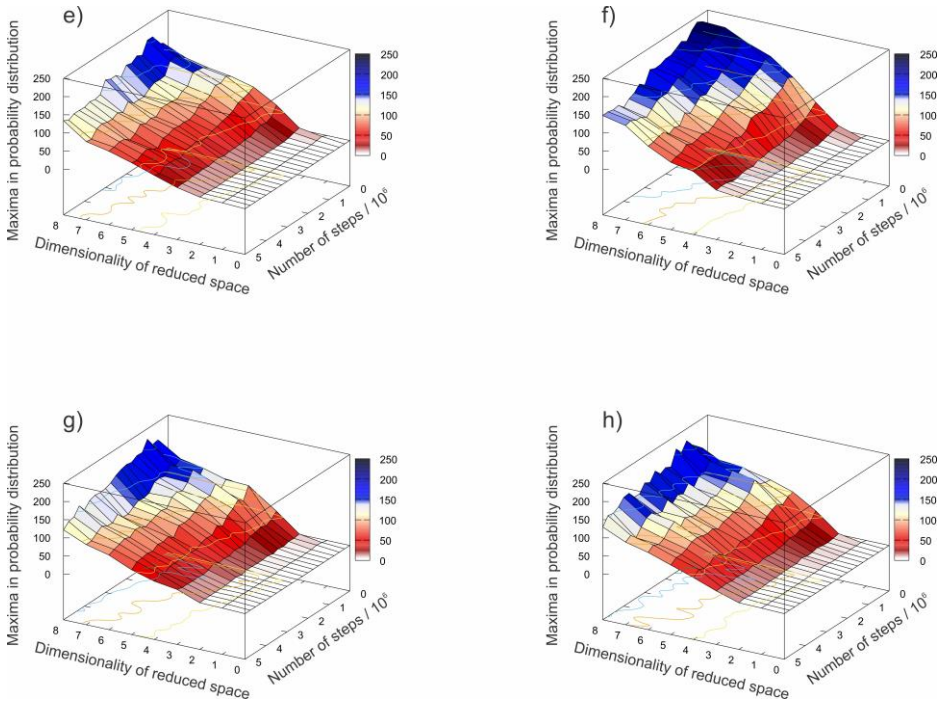


Figure 11. *Plateaus* of strict local maxima in dependence on the total number of linearly independent internal coordinate distances (labels *a*–*s* follow classification given in Table 2, label *t* corresponds to the exact representation and is given for comparison). *(Continuation)*

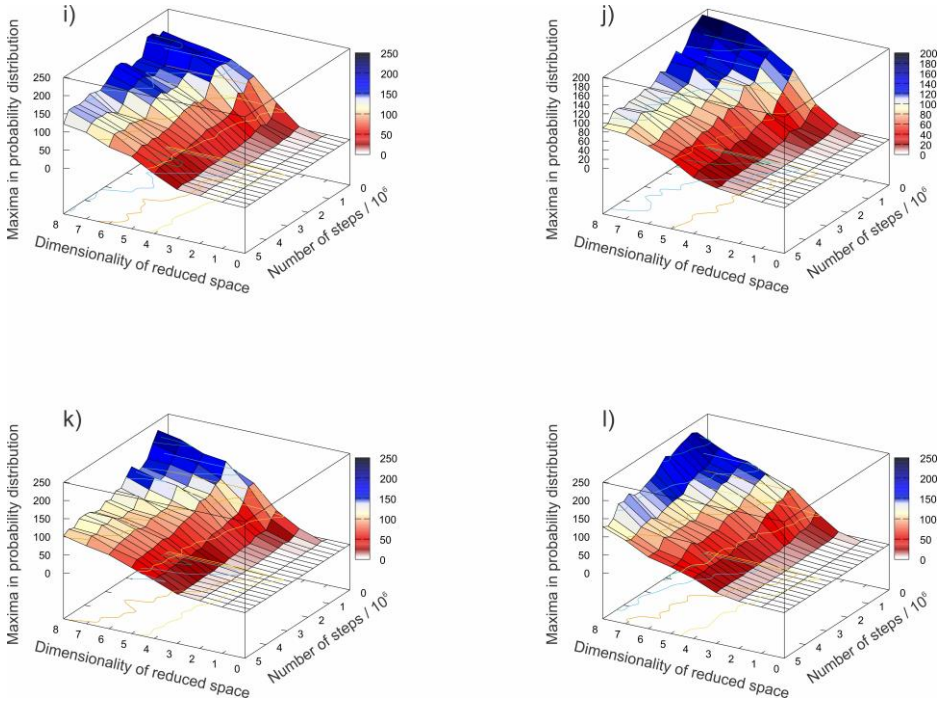


Figure 11. *Plateaus* of strict local maxima in dependence on the total number of linearly independent internal coordinate distances (labels *a*–*s* follow classification given in Table 2, label *t* corresponds to the exact representation and is given for comparison). *(Continuation)*

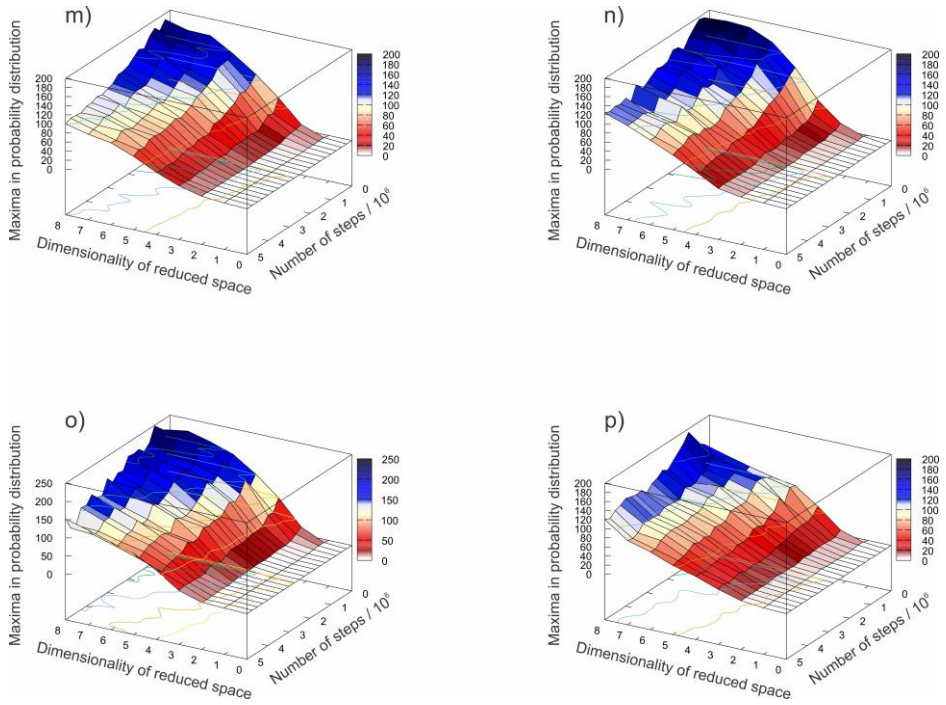


Figure 11. *Plateaus* of strict local maxima in dependence on the total number of linearly independent internal coordinate distances (labels *a*–*s* follow classification given in Table 2, label *t* corresponds to the exact representation and is given for comparison). *(Continuation)*

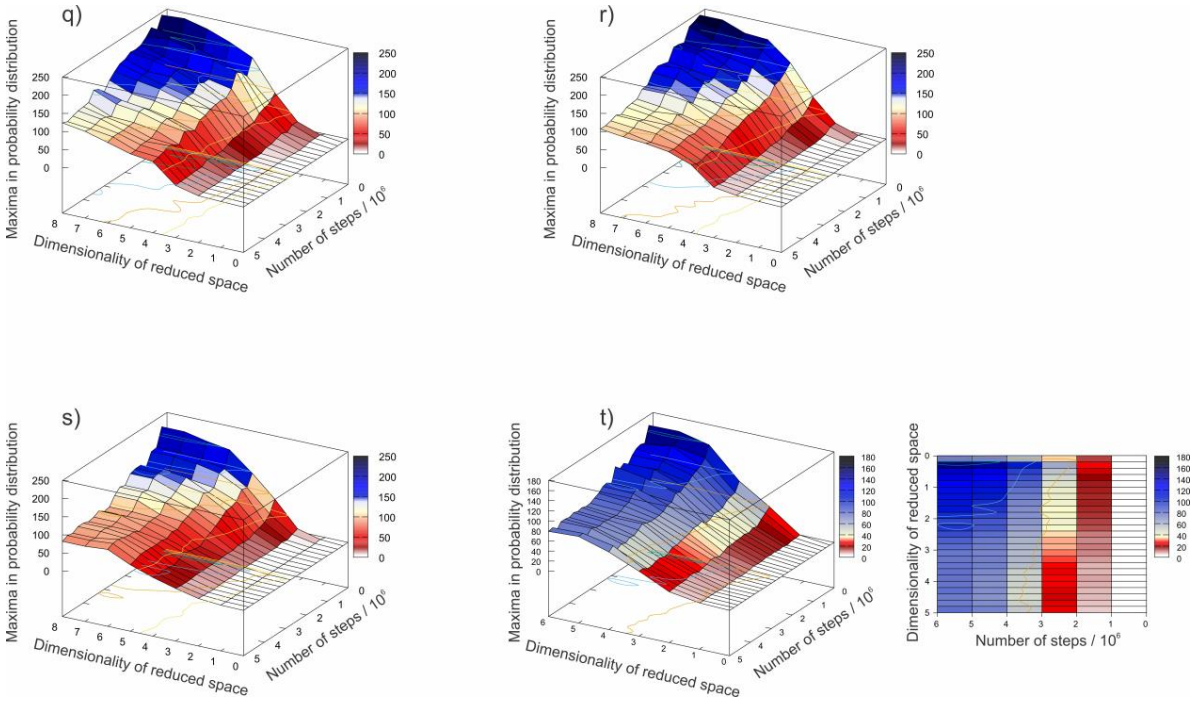


Figure 11. *Plateaus* of strict local maxima in dependence on the total number of linearly independent internal coordinate distances (labels *a*–*s* follow classification given in Table 2, label *t* corresponds to the exact representation and is given for comparison). *(Continuation)*

An increase in the number of distances (and slight changes in definition (see Appendix)) resulted in a fully converged set. These fully converged sets are visible on Figs. 10q–10s. Compared to the exact representation calculated in Cartesian coordinates (Fig. 10t),[6] it is clear that these sets produced the same probability distribution and subsequently the same full conformational space. *Plateau* starts at the dimensionality 6 of the reduced space and approximately 2 000 000 points in the simulation. Just to be on the safe side, slightly longer sampling space is usually used, although all previous calculation shows that the reduced space near the *plateau* already contains information about the full conformational space of the compounds.

§ 6. CONCLUSION

A general procedure for constructing generalized and linearly independent set of internal coordinate distances was established and tested for (*R*)-cinchonidine. This is a molecule of particular interest for us due to the ongoing scientific project and it was already thoroughly investigated. The set was built from a trajectory data obtained by *ab initio* molecular dynamics simulation. Full length of the trajectory was 5 000 000 steps and was computed using *on-the-fly* calculations of forces by PM7 method implemented in MOPAC2016. The temperature was held constant during the simulation using velocity scaling at 1273,15 K.

The most important distances between atom pairs were determined from the points in the trajectory. The length of the trajectory was firstly scanned by 1000 and then by 10 000 points until convergence in the distance coordinates was reached. For each specific length of the trajectory, a machine learning algorithm was applied to eliminate linearly dependent coordinates among all possible defined distances. The algorithm used *leave-one-row-out* method coupled with the rank determination to select those matrix rows (distances) that do not contribute to the overall matrix rank. The optimal representation of distances for different lengths of simulations were determined and tested for convergence by checking the *plateaus* of strict local maxima and conformational space of (*R*)-cinchonidine.

The total number of defined distances in a linearly independent set converged to 220 after the 80 000 points with only slight changes in definition of coordinates compared to the previous sets (70 000 and 60 000). Calculation of strict local maxima *plateaus* confirmed the convergence, and it provided the same results when compared to the exact representation determined in our previous work. Strict local maxima *plateau* was reached using 6 principal components and 3 000 000 points in simulation providing the same probability distribution and the same conformational space as in the exact representation.

§ 7. LIST OF ABBREVIATIONS AND SYMBOLS

BOMD	Born-Oppenheimer Molecular Dynamics
COPD	Chronic Obstructive Pulmonary Disease
DG	Distance Geometry
DFT	Density Functional Theory
IUPAC	International Union of Pure and Applies Chemistry
ML	Machine Learning
MM	Molecular Mechanics
NIPALS	Nonlinear Iterative Partial Least Squares
NMR	Nuclear Magnetic Resonance Spectroscopy
NOESY	Nuclear Overhauser Effect Spectroscopy
PCA	Principal Components Analysis
PC*n*	*n*-th Principal Component
PES	Potential Energy Surface
RMS factor	Root Mean Square factor
SLM	Strict Local Maxima
SVD	Singular Value Decomposition
VS	Verlet – Störmer algorithm

§ 8. REFERENCES

1. D. C. Spellmeyer, A. K. Wong, M. J. Bower, J. M. Blaney, *J. Mol. Graph. Model.* **15** (1997) 18–36.
2. J. Jaramillo-Arango, *Bot. J. Linn. Soc.* **53** (1949) 272–311.
3. J.-P. Starck, L. Provins, B. Christophe, M. Gillard, S. Jadot, P. Lo Brutto, L. Que´re´, P. Talaga, M. Guyaux, *Bioorg. Med. Chem. Lett.* **18** (2008) 2675–2678.
4. G. D. H. Dijkstra, R. M. Kellogg, H. Wynberg, J. S. Svendsen, I. Marko, K. B. Sharpless, *J. Am. Chem. Soc.* **111** (1989) 8069–8076.
5. A. Vargas, A. Baiker, *J. Catal.* 239 (2006) 220–226.
6. K. Sović, T. Ostojić, S. Cepić, A. Ramić, R. Odžak, M. Skočibušić, T. Hrenar, I. Primožič, *Croat. Chem. Acta* **92** (2019) 259-267.
7. MOPAC2016, James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, HTTP://OpenMOPAC.net (2016)
8. T. Hrenar, ***qcc***, *Quantum Chemistry Code*, rev. 0.682, 2020.
9. T. Hrenar, ***moonee***, *Code for Manipulation and Analysis of Multi- and Univariate Data*, rev. 0.6826, 2020.
10. Gaussian 16, Revision C.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G- Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Jr. Montgomery, J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, Gaussian, Inc., Wallingford CT, 2016

11. G. Zamarbide, M. R. Estrada, M. A. Zamora, L. L. Torday, R. D. Enriz, F. T. Vert, G. I. Csizmaida, *Theochem* **666** (2003) 599-608.
12. J. Gasteiger, *Handbook of chemometrics*, Wiley-VCH, Weinheim, 2003, p. 267-269
13. T. Hrenar, I. Primožič, D. Fijan, M. Majerić Elenkov, *Phys. Chem. Chem. Phys.* **19** (2017), 31706-31713.
14. D. Marx, J. Hutter, *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*, Cambridge University Press, London, 2009, p 3-7
15. K. Takatsuka, T. Yonehara, A. Yasuki, K. Hanasaki, *Chemical Theory Beyond Born-Oppenheimer Paradigm*, World Scientific Publishing, Singapore, 2015, p 21-26
16. D. J. Griffiths, D. F. Schroeter, *Introduction to Quantum Mechanics*, Cambridge University Press, United Kingdom, 2018
17. B. T. Sutcliffe, *The Born-Oppenheimer Approximation Methods in Computational Molecular Physics*, 293, Springer, Boston, 1992, p 9-14
18. M. Abramowitz, I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Dover, 1972, p 880
19. A. R. Leach, *Molecular Modelling: Principles and Applications*, Pearson Education Limited, Dorchester, 2nd edn, 2001, p 355-358
20. J. M. Haile, *Molecular Dynamics Simulation: Elementary Methods*, Wiley-Interscience, New York, 1992, p 157-159
21. S-S. Shai, B-D. Shai, *Understanding machine learning: from theory to algorithms*, Cambridge University Press, United Kingdom, 2014, p 1-6
22. T. M. Mitchel, *Machine learning*, McGraw Hill, Burr Ridge, 1997, p. 16-17
23. M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, A. J. Aljaaf: *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*, Springer, United States, 2019, p 3
24. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*, The MIT Press, Cambridge, United Kingdom, 2014, 2015
25. https://data-flair.training/blogs/types-of-machine-learning-algorithms/ (date: 24th December 2020)
26. http://sites.science.oregonstate.edu/math/home/programs/undergrad/CalculusQuestStudyGuides/vcalc/lindep/lindep.html (date: 5th November 2020)

27. https://www.dcs.warwick.ac.uk/people/academic/Steve.Russ/cs131/NOTE11.PDF (date: 5th November 2020)
28. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, The MIT Press, Cambridge, United Kingdom, 2017, p 42-44
29. C. B.Y. Cordella. *PCA: The Basic Building Block of Chemometrics, Analytical Chemistry*, Ira S. Krull, IntechOpen (date: November 7th 2012)
30. J. Shlens, *Int J Rem Sens* **51** (2014)
31. A. Radman Kastelic, R. Odžak, I. Pezdirc, K. Sović, T. Hrenar, A. Čipak Gašparović, M. Skočibušić, I. Primožič, *Molecules* **24** (2019) 2675.

§ 9. APPENDIX

Table A1. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 1000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 1000				
1 2	1 25	9 19	9 40	10 29
1 7	1 32	9 24	9 41	10 31
1 8	1 34	9 25	9 42	10 33
1 9	9 11	9 28	9 43	41 44
1 10	9 12	9 30	9 44	42 43
1 19	9 14	9 32	10 11	42 44
1 21	9 15	9 34	10 20	43 44
1 22	9 16	9 38	10 26	
1 24	9 17	9 39	10 27	

Table A2. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 2000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 2000				
1 2	8 43	13 16	36 44	39 40
1 3	8 44	16 17	37 38	39 44
1 10	9 10	16 19	37 39	40 41
8 34	9 43	16 20	37 40	40 42
8 35	9 44	16 27	37 41	40 43
8 36	10 16	16 38	37 43	40 44
8 37	10 23	36 37	37 44	41 42
8 38	10 25	36 38	38 39	41 43
8 39	10 27	36 40	38 40	42 43
8 40	11 12	36 41	38 41	43 44
8 41	11 15	36 42	38 43	
8 42	13 15	36 43	38 44	

Table A3. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 3000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 3000				
1 3	8 43	10 43	36 37	38 44
1 4	8 44	15 38	36 40	39 40
1 6	9 10	15 39	36 41	39 43
1 9	9 20	15 40	36 42	39 44
1 10	9 23	15 41	36 44	40 41
8 33	9 27	15 42	37 38	40 42
8 34	9 36	15 43	37 39	40 44
8 35	9 40	16 17	37 40	41 42
8 36	9 41	35 38	37 41	41 43
8 37	9 43	35 39	37 43	41 44
8 38	9 44	35 40	37 44	42 43
8 39	10 36	35 41	38 39	43 44
8 40	10 37	35 42	38 40	
8 41	10 41	35 44	38 43	

Table A4. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 4000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 4000				
1 2	8 34	10 34	36 38	38 43
1 7	8 35	10 36	36 39	38 44
1 8	8 36	10 37	36 40	39 40
1 9	8 37	10 41	36 41	39 41
8 18	8 38	10 42	36 42	39 42
8 19	8 39	10 43	36 43	39 43
8 21	8 40	34 44	36 44	39 44
8 22	8 41	35 36	37 38	40 41
8 25	8 42	35 37	37 39	40 43
8 26	8 43	35 38	37 40	40 44
8 28	8 44	35 39	37 41	41 42
8 29	9 22	35 40	37 43	41 43
8 30	9 23	35 41	37 44	41 44
8 31	9 34	35 43	38 39	42 43
8 32	9 35	35 44	38 40	43 44
8 33	9 36	36 37	38 41	

Table A5. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 5000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 5000				
7 42	8 31	9 43	36 39	39 41
7 43	8 32	9 44	36 40	39 42
7 44	8 33	10 27	36 41	39 43
8 10	8 34	10 34	36 43	39 44
8 11	8 35	34 41	36 44	40 41
8 15	8 37	34 43	37 38	40 42
8 16	8 38	34 44	37 39	40 43
8 17	8 39	35 36	37 40	40 44
8 19	8 40	35 37	37 41	41 42
8 21	8 41	35 38	37 43	41 43
8 22	8 42	35 39	37 44	41 44
8 25	8 43	35 40	38 39	42 43
8 26	8 44	35 41	38 40	42 44
8 27	9 23	35 43	38 41	43 44
8 28	9 25	35 44	38 42	
8 29	9 26	36 37	38 44	
8 30	9 27	36 38	39 40	

Table A6. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 6000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 6000				
6 44	8 30	9 43	36 38	38 44
7 43	8 31	9 44	36 39	39 40
7 44	8 32	10 23	36 40	39 41
8 9	8 33	10 28	36 41	39 42
8 13	8 34	10 31	36 42	39 43
8 14	8 35	10 34	36 43	39 44
8 15	8 36	10 41	36 44	40 41
8 18	8 37	10 42	37 38	40 43
8 19	8 38	10 43	37 39	40 44
8 21	8 39	11 44	37 40	41 42
8 22	8 40	35 38	37 41	41 43
8 23	8 41	35 39	37 42	41 44
8 24	8 44	35 40	37 43	42 43
8 25	9 19	35 41	37 44	42 44
8 26	9 20	35 42	38 39	43 44
8 27	9 38	35 43	38 40	
8 28	9 41	35 44	38 41	
8 29	9 42	36 37	38 43	

Table A7. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 7000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 7000				
1 2	8 26	9 38	35 43	38 41
1 5	8 27	9 40	35 44	38 43
6 33	8 28	9 41	36 37	38 44
6 35	8 29	9 42	36 38	39 40
6 44	8 30	9 43	36 39	39 41
7 43	8 31	10 22	36 40	39 42
8 9	8 32	10 23	36 41	39 43
8 11	8 33	10 31	36 42	39 44
8 14	8 35	10 42	36 43	40 41
8 16	8 36	10 43	36 44	40 43
8 17	8 37	10 44	37 38	40 44
8 18	8 38	34 44	37 39	41 42
8 19	8 39	35 36	37 40	41 43
8 20	8 40	35 37	37 41	41 44
8 21	9 26	35 38	37 42	42 43
8 22	9 27	35 39	37 43	42 44
8 23	9 28	35 40	37 44	43 44
8 24	9 32	35 41	38 39	
8 25	9 37	35 42	38 40	

Table A8. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 8000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 8000				
3 21	8 33	33 34	35 39	37 44
6 33	8 34	33 35	35 40	38 39
6 34	8 35	33 36	35 41	38 40
6 35	8 38	33 37	35 42	38 41
6 37	8 39	33 40	35 43	38 43
6 38	8 40	33 41	35 44	38 44
6 39	8 44	33 42	36 37	39 40
6 40	9 23	33 43	36 38	39 42
6 43	9 25	33 44	36 39	39 43
8 12	10 42	34 35	36 40	39 44
8 14	10 43	34 36	36 41	40 41
8 16	32 33	34 37	36 42	40 43
8 17	32 34	34 38	36 43	40 44
8 19	32 36	34 40	36 44	41 42
8 20	32 38	34 41	37 38	41 43
8 23	32 40	34 42	37 39	41 44
8 26	32 41	34 44	37 40	42 43
8 27	32 42	35 36	37 41	42 44
8 28	32 43	35 37	37 42	43 44
8 29	32 44	35 38	37 43	

Table A9. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 9000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 9000				
1 2	30 43	32 42	35 36	37 44
3 5	30 44	32 43	35 37	38 39
29 31	31 32	32 44	35 38	38 40
29 32	31 33	33 34	35 39	38 41
29 33	31 34	33 35	35 40	38 43
29 34	31 35	33 36	35 41	38 44
29 39	31 36	33 37	35 42	39 40
29 40	31 37	33 38	35 43	39 41
29 41	31 38	33 39	35 44	39 42
29 42	31 39	33 40	36 37	39 44
29 43	31 41	33 41	36 38	40 41
29 44	31 42	33 42	36 39	40 42
30 31	31 43	33 43	36 40	40 43
30 32	31 44	33 44	36 41	40 44
30 33	32 33	34 35	36 42	41 42
30 34	32 34	34 36	36 43	41 43
30 35	32 35	34 37	36 44	41 44
30 36	32 36	34 38	37 38	42 43
30 37	32 37	34 39	37 39	42 44
30 39	32 38	34 40	37 40	43 44
30 40	32 39	34 41	37 41	
30 41	32 40	34 43	37 42	
30 42	32 41	34 44	37 43	

Table A10. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 10 000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 10 000				
28 40	30 39	32 39	34 42	37 42
28 41	30 40	32 40	34 43	37 43
28 42	30 41	32 41	34 44	37 44
28 43	30 42	32 42	35 36	38 39
28 44	30 43	32 43	35 37	38 40
29 30	30 44	32 44	35 38	38 42
29 31	31 32	33 34	35 39	38 43
29 32	31 33	33 35	35 40	38 44
29 33	31 34	33 36	35 41	39 40
29 34	31 35	33 37	35 42	39 41
29 38	31 36	33 38	35 43	39 42
29 39	31 37	33 39	35 44	39 44
29 40	31 38	33 40	36 37	40 41
29 41	31 40	33 41	36 38	40 42
29 42	31 41	33 42	36 39	40 43
29 43	31 42	33 43	36 40	40 44
29 44	31 43	33 44	36 41	41 42
30 31	31 44	34 35	36 42	41 43
30 32	32 33	34 36	36 43	41 44
30 33	32 34	34 37	36 44	42 43
30 34	32 35	34 38	37 38	42 44
30 36	32 36	34 39	37 39	43 44
30 37	32 37	34 40	37 40	
30 38	32 38	34 41	37 41	

Table A11. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 20 000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 20 000				
2 41	29 31	31 35	33 44	37 41
2 42	29 32	31 36	34 35	37 42
2 43	29 33	31 37	34 36	37 43
2 44	29 34	31 38	34 37	37 44
3 5	29 35	31 39	34 38	38 39
3 7	29 36	31 40	34 39	38 40
27 33	29 37	31 41	34 40	38 41
27 34	29 38	31 42	34 41	38 42
27 36	29 39	31 43	34 42	38 44
27 38	29 41	31 44	34 43	39 40
27 39	29 42	32 33	34 44	39 41
27 40	29 43	32 34	35 36	39 42
27 42	29 44	32 35	35 37	39 43
27 43	30 31	32 36	35 38	39 44
27 44	30 32	32 37	35 39	40 41
28 29	30 33	32 38	35 40	40 42
28 30	30 34	32 39	35 42	40 43
28 31	30 35	32 40	35 43	40 44
28 32	30 36	32 41	35 44	41 42
28 33	30 37	32 42	36 37	41 43
28 34	30 38	32 43	36 38	41 44
28 37	30 39	32 44	36 39	42 43
28 38	30 40	33 34	36 40	42 44
28 39	30 41	33 35	36 41	43 44
28 40	30 42	33 36	36 42	
28 41	30 43	33 37	36 43	
28 42	30 44	33 40	36 44	
28 43	31 32	33 41	37 38	
28 44	31 33	33 42	37 39	
29 30	31 34	33 43	37 40	

Table A12. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 30 000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 30 000				
1 2	28 34	30 40	33 39	37 38
2 19	28 35	30 41	33 41	37 39
2 21	28 36	30 42	33 42	37 40
2 22	28 37	30 43	33 43	37 41
3 4	28 38	30 44	33 44	37 42
26 38	28 40	31 32	34 35	37 43
26 39	28 41	31 33	34 36	37 44
26 40	28 42	31 34	34 37	38 39
26 41	28 43	31 35	34 38	38 40
26 42	28 44	31 36	34 39	38 41
26 43	29 30	31 38	34 40	38 42
26 44	29 31	31 40	34 41	38 43
27 28	29 32	31 41	34 42	38 44
27 29	29 33	31 42	34 43	39 40
27 30	29 34	31 43	34 44	39 41
27 31	29 35	31 44	35 36	39 43
27 32	29 36	32 33	35 37	39 44
27 33	29 37	32 34	35 38	40 41
27 34	29 39	32 36	35 39	40 42
27 35	29 40	32 37	35 40	40 43
27 36	29 41	32 38	35 41	40 44
27 37	29 42	32 39	35 42	41 42
27 38	29 43	32 40	35 43	41 43
27 40	29 44	32 41	35 44	41 44
27 41	30 31	32 42	36 37	42 43
27 42	30 32	32 43	36 38	42 44
27 44	30 33	32 44	36 39	43 44
28 29	30 34	33 34	36 40	
28 30	30 35	33 35	36 41	
28 31	30 36	33 36	36 42	
28 32	30 37	33 37	36 43	
28 33	30 38	33 38	36 44	

Table A13. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 40 000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 40 000				
1 2	27 43	30 34	33 34	36 43
1 6	27 44	30 35	33 35	36 44
2 9	28 29	30 36	33 36	37 38
2 11	28 30	30 37	33 37	37 39
2 12	28 31	30 38	33 38	37 40
2 13	28 32	30 39	33 40	37 41
2 14	28 33	30 40	33 41	37 42
2 15	28 34	30 41	33 42	37 43
2 17	28 35	30 42	33 43	37 44
2 19	28 36	30 43	33 44	38 39
2 22	28 37	30 44	34 35	38 40
2 40	28 39	31 32	34 36	38 41
2 42	28 40	31 33	34 37	38 42
2 44	28 41	31 34	34 38	38 43
3 4	28 42	31 35	34 39	38 44
3 11	28 43	31 36	34 40	39 40
3 12	28 44	31 37	34 41	39 41
3 13	29 30	31 39	34 42	39 42
3 14	29 31	31 40	34 43	39 43
3 16	29 32	31 41	34 44	39 44
3 17	29 33	31 42	35 36	40 41
3 19	29 34	31 43	35 37	40 42
27 29	29 35	31 44	35 38	40 43
27 30	29 36	32 33	35 39	40 44
27 31	29 37	32 34	35 40	41 42
27 32	29 38	32 36	35 41	41 43
27 33	29 40	32 37	35 42	41 44
27 34	29 41	32 38	35 44	42 43
27 35	29 42	32 39	36 37	42 44
27 36	29 43	32 40	36 38	43 44
27 38	29 44	32 41	36 39	
27 40	30 31	32 42	36 40	
27 41	30 32	32 43	36 41	
27 42	30 33	32 44	36 42	

Table A14. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 50 000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N* (points) = 50 000				
1 2	27 35	30 31	32 41	36 38
1 6	27 36	30 32	32 42	36 39
25 37	27 38	30 33	32 43	36 40
25 38	27 39	30 34	32 44	36 41
25 39	27 41	30 35	33 34	36 42
25 40	27 42	30 36	33 35	36 43
25 41	27 43	30 37	33 36	36 44
25 42	27 44	30 38	33 37	37 38
25 43	28 29	30 39	33 39	37 39
25 44	28 30	30 40	33 40	37 40
26 27	28 31	30 41	33 41	37 41
26 28	28 32	30 42	33 42	37 42
26 29	28 33	30 43	33 43	37 43
26 31	28 34	30 44	33 44	37 44
26 32	28 35	31 32	34 35	38 39
26 33	28 37	31 33	34 36	38 40
26 34	28 38	31 34	34 37	38 41
26 35	28 39	31 35	34 38	38 43
26 36	28 42	31 36	34 39	38 44
26 37	28 43	31 37	34 40	39 40
26 38	28 44	31 38	34 41	39 41
26 39	29 30	31 39	34 42	39 42
26 40	29 31	31 40	34 43	39 43
26 41	29 32	31 41	34 44	39 44
26 42	29 33	31 42	35 36	40 41
26 43	29 34	31 43	35 37	40 42
26 44	29 35	31 44	35 38	40 43
27 28	29 36	32 33	35 39	40 44
27 29	29 37	32 34	35 40	41 42
27 30	29 39	32 35	35 41	41 43
27 31	29 41	32 36	35 42	41 44
27 32	29 42	32 37	35 43	42 43
27 33	29 43	32 38	35 44	42 44
27 34	29 44	32 39	36 37	43 44

Table A15. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 60 000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N* (points) = 60 000									
1 2	25 26	26 32	27 36	28 44	30 37	32 35	34 37	36 41	40 41
1 6	25 28	26 33	27 37	29 30	30 38	32 36	34 38	36 42	40 42
2 27	25 29	26 34	27 38	29 31	30 39	32 37	34 39	36 43	40 43
2 35	25 30	26 35	27 40	29 32	30 40	32 38	34 40	36 44	40 44
2 39	25 31	26 36	27 41	29 33	30 42	32 39	34 41	37 38	41 42
2 43	25 32	26 37	27 42	29 34	30 43	32 41	34 42	37 39	41 43
2 44	25 33	26 38	27 43	29 35	30 44	32 42	34 43	37 40	41 44
3 4	25 34	26 39	27 44	29 36	31 32	32 43	34 44	37 41	42 43
24 27	25 35	26 40	28 29	29 37	31 33	32 44	35 36	37 42	42 44
24 28	25 36	26 41	28 30	29 38	31 34	33 34	35 37	37 43	43 44
24 29	25 37	26 42	28 32	29 39	31 35	33 35	35 38	37 44	
24 30	25 38	26 43	28 33	29 41	31 36	33 36	35 39	38 39	
24 31	25 39	26 44	28 34	29 42	31 37	33 37	35 40	38 40	
24 33	25 41	27 28	28 35	29 43	31 38	33 39	35 41	38 41	
24 34	25 42	27 29	28 36	29 44	31 39	33 40	35 42	38 42	
24 36	25 43	27 30	28 37	30 31	31 40	33 41	35 43	38 44	
24 37	25 44	27 31	28 38	30 32	31 41	33 42	35 44	39 40	
24 38	26 27	27 32	28 39	30 33	31 42	33 43	36 37	39 41	
24 39	26 28	27 33	28 40	30 34	31 43	33 44	36 38	39 42	
24 43	26 29	27 34	28 42	30 35	31 44	34 35	36 39	39 43	
24 44	26 31	27 35	28 43	30 36	32 34	34 36	36 40	39 44	

Table A16. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 70 000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N* (points) = 70 000									
1 2	24 33	25 39	26 43	28 34	29 44	31 40	33 42	36 38	39 42
1 6	24 34	25 40	26 44	28 35	30 31	31 41	33 43	36 39	39 43
2 5	24 35	25 41	27 28	28 36	30 32	31 42	33 44	36 40	39 44
2 6	24 36	25 42	27 29	28 37	30 33	31 43	34 35	36 41	40 41
2 8	24 37	25 43	27 30	28 38	30 34	31 44	34 36	36 42	40 42
2 13	24 38	25 44	27 31	28 40	30 35	32 34	34 37	36 43	40 43
2 14	24 39	26 27	27 32	28 42	30 36	32 35	34 38	36 44	40 44
2 15	24 40	26 28	27 33	28 43	30 37	32 36	34 39	37 38	41 42
2 24	24 43	26 29	27 34	28 44	30 38	32 37	34 40	37 39	41 43
2 29	24 44	26 30	27 36	29 30	30 40	32 39	34 41	37 40	41 44
2 35	25 26	26 31	27 37	29 31	30 41	32 41	34 42	37 41	42 43
2 37	25 27	26 32	27 38	29 32	30 42	32 42	34 43	37 42	42 44
2 38	25 28	26 33	27 39	29 33	30 43	32 43	34 44	37 43	43 44
2 39	25 29	26 34	27 40	29 34	30 44	32 44	35 36	37 44	
2 40	25 31	26 35	27 41	29 35	31 32	33 34	35 37	38 39	
2 44	25 32	26 36	27 42	29 36	31 33	33 35	35 38	38 40	
3 5	25 33	26 37	27 43	29 37	31 34	33 36	35 39	38 41	
3 6	25 34	26 38	27 44	29 38	31 35	33 37	35 40	38 42	
3 7	25 35	26 39	28 29	29 40	31 36	33 38	35 41	38 43	
24 28	25 36	26 40	28 31	29 41	31 37	33 39	35 42	38 44	
24 29	25 37	26 41	28 32	29 42	31 38	33 40	35 44	39 40	
24 30	25 38	26 42	28 33	29 43	31 39	33 41	36 37	39 41	

Table A17. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 80 000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N* (points) = 80 000									
1 2	23 42	25 30	26 36	27 42	29 35	31 33	33 34	35 37	38 39
1 11	23 43	25 31	26 37	27 43	29 36	31 34	33 35	35 38	38 40
1 12	23 44	25 32	26 38	27 44	29 37	31 35	33 36	35 39	38 41
1 13	24 26	25 33	26 39	28 29	29 38	31 36	33 37	35 40	38 42
2 16	24 27	25 34	26 40	28 30	29 40	31 37	33 38	35 41	38 43
2 17	24 29	25 35	26 41	28 31	29 41	31 38	33 39	35 42	38 44
2 21	24 30	25 36	26 42	28 32	29 42	31 39	33 40	35 44	39 40
2 22	24 31	25 37	26 43	28 33	29 43	31 40	33 41	36 37	39 41
2 24	24 32	25 38	26 44	28 34	29 44	31 41	33 42	36 38	39 42
2 27	24 34	25 39	27 28	28 35	30 31	31 42	33 43	36 39	39 43
2 29	24 35	25 40	27 29	28 36	30 32	31 43	33 44	36 40	39 44
2 36	24 36	25 41	27 30	28 37	30 34	31 44	34 35	36 41	40 41
2 37	24 37	25 42	27 32	28 38	30 35	32 33	34 36	36 42	40 42
2 38	24 38	25 43	27 33	28 40	30 36	32 34	34 37	36 43	40 43
2 39	24 39	25 44	27 34	28 41	30 37	32 35	34 38	36 44	40 44
2 40	24 42	26 27	27 35	28 43	30 38	32 36	34 39	37 38	41 42
2 42	24 43	26 28	27 36	28 44	30 39	32 37	34 40	37 39	41 43
2 44	24 44	26 29	27 37	29 30	30 40	32 40	34 41	37 40	41 44
3 4	25 26	26 30	27 38	29 31	30 42	32 41	34 42	37 41	42 43
3 5	25 27	26 33	27 39	29 32	30 43	32 42	34 43	37 42	42 44
3 7	25 28	26 34	27 40	29 33	30 44	32 43	34 44	37 43	43 44
23 41	25 29	26 35	27 41	29 34	31 32	32 44	35 36	37 44	

Table A18. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 90 000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N* (points) = 90 000									
1 2	23 41	25 29	26 35	27 42	29 36	30 44	32 43	35 36	37 44
1 11	23 42	25 30	26 36	27 43	29 37	31 32	32 44	35 37	38 39
1 12	23 43	25 31	26 37	27 44	29 38	31 33	33 34	35 38	38 40
1 35	23 44	25 32	26 38	28 29	29 39	31 34	33 35	35 39	38 41
2 16	24 26	25 33	26 40	28 30	29 40	31 35	33 36	35 40	38 42
2 17	24 27	25 34	26 41	28 31	29 41	31 36	33 37	35 41	38 43
2 20	24 28	25 35	26 42	28 32	29 42	31 37	33 38	35 43	38 44
2 21	24 29	25 36	26 43	28 33	29 43	31 38	33 40	35 44	39 40
2 22	24 30	25 37	26 44	28 34	29 44	31 39	33 41	36 37	39 41
2 24	24 31	25 38	27 28	28 35	30 31	31 40	33 42	36 38	39 42
2 27	24 32	25 39	27 29	28 36	30 32	31 41	33 43	36 39	39 43
2 29	24 34	25 41	27 30	28 37	30 33	31 42	33 44	36 40	39 44
2 36	24 35	25 42	27 32	28 38	30 34	31 43	34 35	36 41	40 41
2 37	24 36	25 43	27 33	28 39	30 35	31 44	34 36	36 42	40 42
2 38	24 37	25 44	27 34	28 40	30 36	32 34	34 37	36 43	40 43
2 39	24 38	26 27	27 35	28 41	30 37	32 35	34 38	36 44	40 44
2 42	24 39	26 28	27 36	28 44	30 38	32 36	34 39	37 38	41 42
2 44	24 43	26 29	27 37	29 31	30 39	32 37	34 40	37 39	41 43
3 4	24 44	26 30	27 38	29 32	30 40	32 39	34 41	37 40	41 44
3 5	25 26	26 32	27 39	29 33	30 41	32 40	34 42	37 41	42 43
3 7	25 27	26 33	27 40	29 34	30 42	32 41	34 43	37 42	42 44
3 10	25 28	26 34	27 41	29 35	30 43	32 42	34 44	37 43	43 44

Table A19. The set of linearly independent generalized coordinates as a result of machine learning algorithms performed on trajectory with 100 000 points for a (*R*)-cinchonidine molecule.

The set of linearly independent generalized coordinates for trajectory with *N*(points) = 100 000									
1 2	23 41	25 28	26 34	27 42	29 34	31 32	32 42	35 36	37 44
2 16	23 42	25 29	26 35	27 43	29 35	31 33	32 43	35 37	38 39
2 17	23 43	25 30	26 36	27 44	29 36	31 34	32 44	35 38	38 40
2 20	23 44	25 31	26 37	28 29	29 37	31 35	33 34	35 39	38 41
2 21	24 26	25 32	26 38	28 30	29 38	31 36	33 35	35 40	38 42
2 22	24 27	25 33	26 39	28 31	29 39	31 37	33 36	35 41	38 43
2 24	24 29	25 34	26 41	28 32	29 40	31 38	33 37	35 43	38 44
2 27	24 30	25 35	26 42	28 33	29 41	31 39	33 40	35 44	39 40
2 29	24 31	25 36	26 43	28 34	29 42	31 40	33 41	36 37	39 41
2 36	24 32	25 37	26 44	28 35	29 43	31 41	33 42	36 38	39 42
2 37	24 34	25 38	27 28	28 36	29 44	31 42	33 43	36 39	39 43
2 38	24 35	25 39	27 29	28 37	30 31	31 43	33 44	36 40	39 44
2 39	24 36	25 40	27 30	28 38	30 32	31 44	34 35	36 41	40 41
2 40	24 37	25 41	27 32	28 39	30 34	32 33	34 36	36 42	40 42
2 44	24 38	25 42	27 33	28 40	30 35	32 34	34 37	36 43	40 43
3 4	24 39	25 43	27 34	28 41	30 36	32 35	34 38	36 44	40 44
3 5	24 40	25 44	27 35	28 43	30 37	32 36	34 39	37 38	41 42
3 7	24 42	26 27	27 36	28 44	30 38	32 37	34 40	37 39	41 43
3 10	24 43	26 28	27 37	29 30	30 39	32 38	34 41	37 40	41 44
23 38	24 44	26 29	27 38	29 31	30 42	32 39	34 42	37 41	42 43
23 39	25 26	26 30	27 39	29 32	30 43	32 40	34 43	37 42	42 44
23 40	25 27	26 33	27 41	29 33	30 44	32 41	34 44	37 43	43 44

§ 10. CURRICULUM VITAE

Personal Information

Name and surname: Tea Ostojić

Date of birth: 15th March 1995

Place of birth: Zadar, Republic of Croatia

Education

2001–2002	Primary school "*Šime Budinić*", Zadar
2002–2009	Primary school "*Antun Mihanović*", Zagreb
2009–2013	High school "*3. gimnazija*", Zagreb
2013–2018	Undergraduate study in Chemistry, Faculty of Science, University of Zagreb, Croatia
2019–2020	Erasmus+ international student exchange, Maria Curie-Skłodowska University, Lublin, Poland
2018–today	Graduate study in Chemistry, fields: Inorganic and Physical chemistry, Faculty of Science, University of Zagreb, Croatia

Honours and Awards

2019 Award for student's scientific research in academic year 2018/2019, Department of Chemistry (Faculty of Science)

Activities in Popularization of Science

2013–2019	member of PRIMUS (student's association)
	Dan i noć na PMF-u

Scientific Publications

1. K. Sović, **T. Ostojić**, S. Cepić, A. Ramić, R. Odžak, M. Skočibušić, T. Hrenar, I. Primožič, *Croat. Chem. Acta*, **92** (2019), 259-267.

§ 11. IZJAVA O LEKTURI

IZJAVA O LEKTURI

Ja, Veronika Mišura, dipl. angl. i germ. i sudski tumač za engleski i njemački jezik, izjavljujem da je rad naslova:

MACHINE LEARNING ASSISTED DETERMINATION OF LINEARLY INDEPENDENT SET OF GENERALIZED MOLECULAR COORDINATES

autorice Tee Ostojić lektoriran prema pravilima hrvatskoga odnosno engleskoga jezika.

Datum

17. svibnja 2021.

Potpis lektora

STALNI SUDSKI TUMAČ ZA ENGLESKI I NJEMAČKI JEZIK
Court interpreter - Ständige Gerichtsdolmetscherin
Autorizacija
VERONIKA MIŠURA
ZAGREB-SESVETE
Selnička 10
tel 01/2005 189
fax 01/2005 333